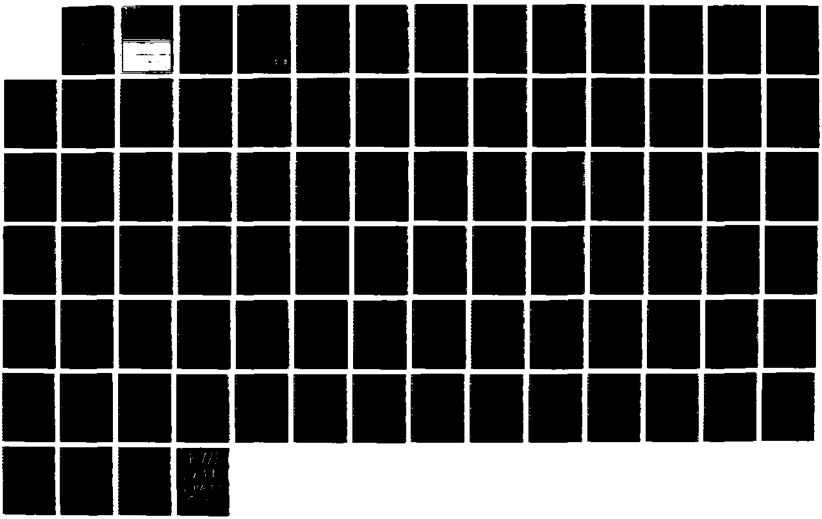
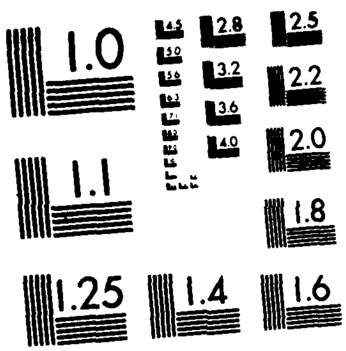
AD-A191 050 ADAPTIVE TIME SERIES ANALYSIS USING PREDICTIVE INFERENCE AND ENTROPY(U) SCIENTIFIC SYSTEMS INC CAMBRIDGE MA D E DUSTAFSON DEC 87 AFOSR-TR-88-0032 1/1

UNCLASSIFIED F9620-87-C-0026 F/G 12/1 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

DTIC FILE COPY

(2)

AFOSR-TR- 88-0032

AD-A191 858

Scientific Systems

88 2 24 092

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
...roved for public release,
distribution unlimited

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 88-0032

6a. NAME OF PERFORMING ORGANIZATION
Scientific Systems, Inc.

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
DCASMA Boston

6c. ADDRESS (City, State, and ZIP Code)
One Alewife Place
Cambridge, MA 02140

7b. ADDRESS (City, State, and ZIP Code)
495 Summer Street
Boston, MA 02210-2184

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
USAF Office of Scient. Res.

8b. OFFICE SYMBOL (If applicable)
NM

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
F49620-87-C-0026

8c. ADDRESS (City, State, and ZIP Code)
Building 410
Bolling AFB, DC 20332-6448

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO
61102F	2305	A1	

11. TITLE (Include Security Classification)
Adaptive Time Series Analysis Using Predictive Inference And Entropy

12. PERSONAL AUTHOR(S)
Donald E. Gustafson

13a. TYPE OF REPORT
Annual

13b. TIME COVERED
FROM 12/86 TO 12/87

14. DATE OF REPORT (Year, Month, Day)
88/1/15

15. PAGE COUNT

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Research is being conducted on adaptive time series methods for detecting and tracking both abrupt and slow changes in both structure and parameters. The methods are based on a unified statistical frame work which is motivated by statistical inference and entropy arguments. The method yields estimates of input/output dynamics and noise statistics. An integrated approach which combines canonical variates analysis and maximum-likelihood estimation has been developed and tested. Specific attention is given to the problem of parameter truncation in both a linear predictor and Kalman filter framework.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Maj Woodruff

22b. TELEPHONE (Include Area Code)
202-767-5027

22c. OFFICE SYMBOL
NM

DD Form 1473, JUN 86 Previous editions are obsolete. SECURITY CLASSIFICATION OF THIS PAGE

(2)

Annual Technical Report

ADAPTIVE TIME SERIES ANALYSIS
USING PREDICTIVE INFERENCE AND ENTROPY

December 1987

By:

Donald E. Dustafson
SCIENTIFIC SYSTEMS, INC.
One Alewife Place
Cambridge, MA 02140

Prepared for:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Bolling Air Force Base
Washington, DC 20332-6448

Under Contract No. F49620-87-C-0026

CONTENTS

CONTENTS

Figures | *Page*

CONTENTS

Tables | **Page**

1. INTRODUCTION

1.1 Overview of the Adaptive Time Series Analysis Problem

Adaptation in time series is an important problem in a number of DOD systems and has many applications in various commercial industries. This is an especially difficult problem in problems requiring realtime adaptation to process changes since such a procedure would have to be completely automatic and reliable. Adaptation is necessary in systems where the dynamical characteristics change with time in unpredictable ways, or where the noise disturbance process characteristics vary with time. Examples of systems that require adaptive time series analysis are the adaptive suppression of aircraft wing flutter, identification of the dynamics of large flexible space structures, detection of failures in aircraft from subsystem failures of battle damage, identification of missile aerodynamics, target tracking, and various signal processing problems.

The solution to the adaptive time series analysis requires several advances in current time series methods. At the core of the problem is the need for a fundamental statistical approach to the adaptation problem that poses the problem in a meaningful way and that leads to computable solutions. To solve the online adaptation problem, a reliable and automatic time series modeling procedure is required that is lacking in previous methods. The current research provides

- A sound statistical basis for posing and solving the adaptation problem

- A numerically and statistically reliable online computational procedure

This approach has been used in conjunction with a new high resolution system identification method utilizing canonical variate analysis (CVA) for the determination of the dynamics of high order multisensor systems with a small data length (Larimore, 1983b). This algorithm can be implemented on highly parallel processors such as a systolic array. This makes practical the consideration of many different system characteristics to determine the best for modeling the observed sensor data and correlational relationships between the many sensors. The system characteristics that have been successfully determined adaptively are the dynamical state order of the system, the presence of correlated disturbances, the optimal data length to use in tracking a time varying system, and the optimal data interval for detection of an abrupt change or other event in the data.

The CVA time series analysis method has been applied to the design of an adaptive flutter suppression problem for suppressing wing flutter or aero-structural vibration in aircraft. While considerable progress has been made in the problem of adaptation in terms of identification of time series models, adaptive time series methods which can efficiently track and detect time varying processes would further improve the system. In such a system the wing dynamical characteristics can change instantaneously when a wing store is dropped, and the new wing dynamics are unknown and may be unstable resulting in a growing oscillation. If the unstable mode is not detected, accurately identified, and stabilized by control feedback in less

than a second, then the aircraft can lose a wing. The CVA algorithm using entropy methods for deciding model state order are being implemented on a vector array processor which will identify high order systems with dozens of dynamical states and multiple inputs and outputs in fractions of a second. This system has been tested in real-time simulations, and was successfully demonstrated in wind tunnel tests at the NASA Langley Transonic Dynamics Wind Tunnel. It is expected that highly parallel processors such as systolic array processors could result in a speedup of many thousands of times which would be required for some very large scale real time adaptive problems.

1.2 Signal and Fault Detection

A Comprehensive survey of fault detection methods is given by Willsky (1976). See also Mehra and Peschon (1971), Willsky and Jones (1974), Willsky (1980), and Isermann (1984). The type of abrupt changes in a system that are considered are of the form

$$x(t+1) = \Phi x(t) + Gu(t) + w(t) + m(t) \tag{1.1}$$

$$y(t) = Hx(t) + Au(t) + Bw(t) + v(t) + N(t) \tag{1.2}$$

where u is the input vector process, y is the output vector, x is the state vector, and w and v are white noise processes that are independent with covariance matrices Q and R respectively. These white noise processes model the covariance structure of the error in predicting y from u. The abrupt changes are in the form of the time the functions m(t) and n(t) introduced into the state and observation equations. Fault detection is thus the detection of the presence of such nonzero functions.

For various hypothesized forms of the functions, i.e., for jumps in various components or specific combinations of the components, a particular detection computation is devised which requires implementation of a Kalman filter. This leads to statistically most powerful likelihood ratio tests of the various failure hypotheses. An optimal solution to the failure detection problem formulated in (1.1) and (1.2) is thus obtained.

There are however several more general failure detection problems not of the form of (1.1) and (1.2). The approach permits only the consideration of simple hypotheses, i.e., where the failure functions m(t) and n(t) are of the form of an unknown scalar amplitude parameter multiplying a function of known form. More general functional forms such as two components with different unknown amplitude parameters multiplying the known functions requires maximum likelihood parameter identification at considerable computational expense and loss of numerical reliability. Furthermore, the problem of unknown failure time leads to a considerable increase in the required computation, and no theoretically sound decision procedure has been proposed for choosing the failure time.

The general case of changes in the system dynamics or correlational characteristics of the disturbance or measurement noise processes cannot be handled. Such cases require general time series analysis parameter identification methods which are not reliable for online application to high state order multivariable systems as discussed in Section Multisensor System Identification. Isermann (1984) gives a survey of current fault detection methods and concludes that: "A unique calculation of the process

coefficients and a parameter estimation with high precision is only possible for low order elements between measured variables. Therefore the measured variables should be selected such that the process is divided in first order elements or, in other words, all state variables should be measurable. Easy to implement parameter estimation methods for continuous-time modles to be used on-line, real-time and in closed loop need to be developed." The requirement of measuring all of the states is not realistic in most situations especially in general multivariate time series and system identification problems. Fortunately, the CVA system identification method does not require this, but indeed is an online, real-time method that gives the same accuracy in either open or closed loop.

The issues of adaptation are not addressed in the fault detection literaure except in simplistic ways. The present state of the art in adaptation for failure detection appears to be the work of Hagglund (1983) discussed in the next section, and is just beginning of adaptive approaches which consider fundamental issues in adaptation.

1.3 Adaptation to Changing Processes

Concepts of adaptive systems have been around since the 1950's involving various senses of adaptation. The present literature on the subject includes a number of methods such as recursive computational schemes, exponential forgetting, lattice computational methods, etc., which have certain "knobs" that allow tuning of the algorithm to accommodate changes in the characteristics of the actual processes. Reviews of these and related methods are contained in several recent special issues of technical

journals and books (Special Issue on Adaptive Control, Automatica, Vol. 20, No. 5, 1985; Special Issue on Linear Adaptive Filtering, IEEE Trans. on Information Theory, Vol. 30, No 2, 1984; Honing and Messerschmitt, 1984). While these methods do permit some degree of adaptation to process changes, the methods of adaptation are ad hoc, and no sound underlying statistical principle for adaptation is proposed or demonstrated. As might be expected, these methods can work poorly on certain cases because of the lack of a sound statistical basis.

In particular, the recursive prediction error and lattice methods are convenient due to their recursive form and provide an estimate at every observation (Friedlander, 1982a, 1982b, 1983; Ljung and Soderstrom, 1983). Also, the recursive algorithms can be sued for adaptation by exponential weighting of the past data (Wellstead and Sanoff, 1981; Irving, 1979; Evans and Betz, 1982). But the rational for exponential weighting has not been given a sound fundamental justification, but is used largely due to its ease of use. the choice of the exponential weight has been ad hoc and susceptible to misinterpretation of changing noise variance levels as time varying changes in the dynamics (Hagglund, 1983).

The fundamental problem in adaptive time series analysis is adaptation to time varying processes. The essential problem is the determination of the characteristics describing the rate at which the process is changing. This problem has received very little in-depth treatment in the literature. Most of the difficulty can be attributed to the discrepancy between the true and assumed uncertainty in the measurements. Adaptive control schemes

are notoriously optimistic about the quality of the parameter estimates because the time varying nature of the process is ignored.

A notable exception is the recent work of Hagglund (1983) which takes an information handling point-of-view. This approach leads to a more realistic appraisal of the accuracy of the parameter estimates and consequently the value of new measurements which become available in time. Two classes of time varying systems are considered:

- Processes with abrupt changes

- Processes with slowly varying changes.

Within each of these classes, changes are considered in the process dynamics and/or noise variance.

For abrupt changes, the fault detection approach is taken. The central idea is to monitor differential changes in the parameter estimates to detect abrupt changes. A new procedure is derived by Hagglund which requires no apriori information and is very sensitive to jumps in the parameters. This procedure is shown to have very good properties in both theory and practice. This works well for parameters of the dynamics as well as those of the noise variances in the simple cases of low order systems.

The problem of slowly varying parameters has plagued many adaptive control schemes. Although the concept of discounting the old data using a forgetting factor has been in use for a long time, the problem of how to

relate this factor to the data has been elusive. The principal proposed by Hagglund is to discount past data in such a way that a constant amount of information would be retained if the parameters were constant. The quantitative measure of the information used is the inverse of the parameter estimation error covariance matrix which is the Fisher information matrix. Theory and simulations show that this works quite well in low order and well conditioned systems. However for high order and multisensor systems with illconditioned parametric structure, the algorithms are not so well behaved.

1.4 Multisensor System Identification

System parameter identification from observed measurements is a crucial part of the adaptive multivariate timeseries analysis problem. It is necessary to adapt anot only to changes in the input to output characteristics of a system, but the correlational characteristics of the disturbance and noise processes must simultaneously be determined. The feasibility of adaptive methods requires first that a reliable online multivariate time series identification procedure be available.

There are several difficulties with currently available methods and software for the identification of system dynamics and noise characteristics. Current methods include the self tuning regulator (STR) (Ljung, 1983; Astrom, 1973; Astrom et al, 1973, 1977), maximum likelihood estimation (MLE) (Mehra and Tyler, 1973; Larimore, 1981a), Box-Jenkins (BJ) methods (Box and Jenkins, 1976), and a variety of heuristic approaches. The current state of the art in both MLE and BJ require that an analyst be

involved in the procedure, and the required number of computational iterations is not bounded. The STR has been applied successfully to simple processes, but is not completely reliable for general processes particularly when multi-input, multi-ouput systems are involved. In addition, the recursive prediction error algorithm used in the STR requires a good initial estimate and so is not suitable for short data where no apriori data is available. The heuristic approaches tend to be special purposes and are rather unreliable in general applications.

Of the current approaches to multivariate time series identification which are high resolution, i.e., make efficient use of the observational information, most use the ARMA (autoregressive moving average) representation for the process. For multi-input multi-output systems this is not a globally well defined parameterization which is a major cause of the difficulties in the present identification methods (Gevers and Wertz, 1982). A consequence is that there is no single parameterization which is numerically well conditioned, and known algorithms can be made to fail for a particular choice of system. The system identification problem is well defined in that the class of models does have best models in a maximum likelihood sense (Larimore, 1981a), but the ARMA parameterization is not unique so that for cases such as pole-zero cancellation there is a whole equivalence class of models with equivalent characteristics. In the sequel this difficulty in parameterization will be resolved by the use of state space models, and stable numerical methods will be described for statistically reliable online identification of multivariable time series.

1.5 Adaptive Time Series Analysis Using Predictive Inference and Entropy

Recently a very general predictive inference approach to statistical modeling has led to a fundamental statistical inference justication of negative entropy as the natural measure of model approximation error (Larimore, 1983a). This development has a number of very attractive features:

- It applies to completely general modeling problems including nonparametric methods.
- It applies exactly to small samples.
- Only the fundamental statistical principlaes of sufficiency and repeated sampling are used.
- It applies to time correlated problems such as time series model identification and tracking.
- Statistical inference can be fundamentally viewed as model approximation.

Early developments in predictive distributions are very old, although modern approaches apparently begin with Jeffreys (1961, p.143) who used a Bayesian approach, as has much of the work following (Atchison and Dunsmore, 1975, preface and p. 39). The approach taken here has been stimulated by Murray (1977, 1979), the work of Akaike (1973) and model structure determination problems (Larimore, 1977a).

1.6 Initial Results Indicating Feasibility

SSI has been in the forefront in developing the CVA and entropy methods. Here the related projects are discussed along with preliminary results indicating the feasibility of the proposed methods.

The original stochastic realization method of Akaike's (1975) was further developed into a commercial software package for mainframe and mini computers by Mehra (1978) and Mehra and Cameron (1976, 1980). Further generalizations to input output systems along with refinements in computational speed and accuracy were developed by larimore (1983b) and Goodrich and Larimore (1983) leading to the current timeseries analysis and forecasting package, Forecast Master (Trademark of SSI), for the IBM/PC. This package is in widespread use in utilities, banks companies and universities.

This algorithm has been the basis for several studies in online systems identification. The project "Basic Research in Adaptive Model Algorithmic Control" used the online CVA system identification algorithm. In the current study "Reconfiguration Control Strategies", the CVA method along with adaptive tracking and detection methods are being studied. The present theory on adaptation using entropy methods (Larimore, 1985a) was developed under the basic research study "traget Dynamic Modeling" and under the study "Development of Statistical Methods Using Predictive Inference and Entropy" which was Phase I of this proposed Phase II study.

A review of the technology in system identification and adaptive control for adaptive methods applicable to the suppression of aeroelastic

wing vibration (flutter) was done in Larimore and Mehra (1984). This study describes the deficiencies of current methods and suggets the feasibility of CVA and entropy methods for fully adaptive online detection and tracking of wing flutter. In a current study with General Dynamics sponsored by the Air Force Wright Aeronautical Laboratories, CVA has been analyzed extensively in computer simulations, real time tests, and demonstrated wind tunnel teste for adaptive flutter suppression. The ability of CVA to identify very complex flutter dynamics of high state order involving very closely spaced spectral peaks in the presence of correlated wind gust disturbances using short data lenghts demonstrated the consderable statistical accuracy of the method. The online CVA identification algorithm was demonstrated ina wind tunnel test at the NASA Langley Transonic Dynamics Wind Tunnel on a 1/4 scale model of an F-16 aircraft.

1.7 Synopsis of Report

In Section 2, we present a detailed and transparent derivation of an unbiased entropy measure which will be used in the sequel for adaptive estimation. This measure is asymptotically equal to Akaike's AIC criterion. In Section 3, we present a detailed description and derivation of linear least-squares prediction using canonical variates analysis (CVA). Several new forms for these predictors are given. In Section 4, a method for direct determination of the parameters of the Kalman filter in canonical form is given, and is shown to be equivalent to a truncated optimal linear predictor derived using CVA. Section 5 considers the model order selection problem, using an entropy-based approach. The problem of abrupt

change detection using entropy methods is considered in Section 6 and a specific algorithm is derived and tested. In Section 7 we consider the problem of slow change detection, specifically the problem of finding the optimal data length for model fitting when the time series coefficients are slowly varying. An entropy-based algorithm is developed and tested.

2. PREDICTIVE INFERENCE AND ENTROPY

2.1 Introduction

In this section we develop the necessary background for development of adaptive estimation algorithms in the sequel.

The problem under consideration is that of predicting the future evolution of a time series, given some observations of the past. The predictive inference framework may be described as follows.

We assume that the density function of interest is parametrized by a parameter vector $\Theta \in R^m$ and is denoted by $p(x | \Theta)$. For the purposes of discrimination between two alternatives Θ_1 and Θ_0 it can be shown (Akaike, 1973) that all necessary information is contained in the likelihood ratio

$$L(x) = \frac{p(x | \Theta_1)}{p(x | \Theta_0)} \tag{2.1}$$

Thus, the mean amount of information for discrimination when $p(x | \Theta_0)$ is the true density is of the form

$$I(\Theta_1, \Theta_0) = \int p(x | \Theta_0) \, \phi \left[\frac{p(x | \Theta_1)}{p(x | \Theta_0)} \right] dx \tag{2.2}$$

where $\phi(.)$ is a properly chosen function. It can be argued using information theoretic arguments (Akaike, 1973) that the only appropriate form is

$$\phi(y) = \log y \tag{2.3}$$

which leads directly to the measure

$$B(\Theta_1, \Theta_0) = \int p(x \mid \Theta_0) \log \left[\frac{p(x \mid \Theta_1)}{p(x)\Theta_0} \right] dx \qquad (2.4)$$

Note that $- B(\Theta_1, \Theta_0)$ is the Kullback-Liebler information for discrimination in favor of Θ_0. It can be easily shown that $B(\Theta_1, \Theta_0) \leq 0$ and equality holds if and only if $p(x \mid \Theta_1) = p(x \mid \Theta_0)$ almost everywhere (Aitchison and Dunsmore, 1975).

Note that $B(\Theta_1, \Theta_0)$ can be written as

$$B(\Theta_1, \Theta_0) = \int p(x \mid \Theta_0) \log p(x \mid \Theta_1)\, dx$$
$$- \int p(x \mid \Theta_0) \log p(x \mid \Theta_0)\, dx \qquad (2.5)$$

Since Θ_0 represents the true (unknown) parameter, our objective is to find the parameter estimate $\hat{\Theta}$ which maximize $B(\hat{\Theta}, \Theta_0)$. From (2.5), we need only maximize

$$\int p(x \mid \Theta_0) \log p(x \mid \hat{\Theta})\, dx$$

with respect to $\hat{\Theta}$ to produce our estimate. This estimate maximizes the expected log-likelihood and is thus a maximum - likelihood estimate.

2.2 Preliminaries

In order to present a clear development, we will work in a partitioned sample space. The random variable x is presumed to be in n - dimensional Euclidean space, $x \in R^n$, and R^n is partitioned into s mutually disjoint regions $\Omega_1, \Omega_2, \ldots, \Omega_s$ which cover R^n:

$$\Omega_1 \cup \Omega_2 \cup \ldots \cup \Omega_s = R^n$$

$$\Omega_i \cap \Omega_j = 0 \; ; \; i \neq j$$

We then define

$$p_i(\Theta) = \int_{\Omega_i} p(x \mid \Theta)\, dx \qquad (2.6)$$

$$i = 1, 2, \ldots, s$$

We consider two different samples, an informative sample q and a predictive sample r. The informative sample is

$$x_q = \{x_{q1}, x_{q2}, \ldots, x_{qn_q}\}$$

which consists of n_q observations of x. The predictive sample is

$$x_r = \{x_{r1}, x_{r2}, \ldots, x_{rn_r}\}$$

consists of n_r observations of x. We assume that n_{qi} of the informative samples fall into Ω_i and that n_{ri} of the predictive sample fall into Ω_i. Then

$$\sum_{i=1}^{s} n_{qi} = n_q$$

$$(2.7)$$

$$\sum_{i=1}^{s} n_{ri} = n_r$$

The two samples x_q and x_r are from the true distribution.

Thus we have, approximately, for sufficiently large samples,

$$p_{qi}(\Theta_0) = \frac{n_{qi}}{n_q} \tag{2.8}$$

and

$$p_{ri}(\Theta_0) = \frac{n_{ri}}{n_r} \tag{2.9}$$

and we assume regularity conditions throughout such that

$$p_q(x \mid \Theta_0) = \lim_{\substack{n_q \to \infty \\ s \to \infty}} p_{qi}(\Theta_0)$$

where

$$\lim_{\substack{s \to \infty \\ x \in \Omega_i}} \Omega_i = x$$

and similarly for $p_r(x \mid \Theta_0)$. The computation of the probabilities associated with the parametrized densities is different. Here we use the definition (2.6) and note that $p_i(\Theta)$ is computable from $p(x \mid \Theta)$ and knowledge of Ω_i. In practice, this computation need not be done, as become clear in the sequel.

2.3 Entropy and Maximum Likelihood Estimation

The first step in our development is to form the maximum-likelihood estimate. This is done by maximuzing (2.5) on the informative sample:

$$\hat{\Theta} = \arg\max_{\Theta} B_q(\Theta, \Theta_0)$$

where

$$B_q(\Theta, \Theta_0) = \sum_{i=1}^{s} p_{qi}(\Theta_0) \log \left[\frac{p_{qi}(\Theta)}{p_{qi}(\Theta_0)} \right] \tag{2.10}$$

Thus

$$\sum_{i=1}^{s} p_{qi}(\Theta_0) \left. \frac{\partial \log p_{qi}(\Theta)}{\partial \Theta} \right|_{\hat{\Theta}} = 0 \tag{2.11}$$

We note here that an approximation for $B_q(\Theta_1, \Theta_0)$ is

$$B_q(\Theta, \Theta_0) \approx \sum_{i=1}^{n_q} \log \frac{p(x_i \mid \Theta)}{p(x_i \mid \Theta_0)} \tag{2.12}$$

and the two expressions are asymptotically equal as $n_q \to \infty$. This form was used by Akaike (1973) to derive the AIC criterion.

Solving (2.12) would, in principal, give the maximum-likelihood estimate <u>if the dimension of Θ were known</u>. However, in practice, the actual dimension, m, of θ is not known. Furthermore, there is an obvious tradeoff between the dimension of our estimate $\hat{\Theta}$ and prediction error. Assume $\hat{\Theta} \in R^k$. Then as we increase k, the fit error on the informative sample will decrease momotonically. However, at some point we are in danger of overfitting the model so that $\hat{\Theta}$ is a function of the sampling error on the informative sample. When this happens, the fit errors on the predictive sample will begin to increase.

If we assume that the true parameter vector dimension is m and that the estimated parameter dimension is $k < m$, then our objective is to evaluate the information measure on the predictive sample and select the model which

maximizes this measure. The discrimination measure is now separated into two parts in order to simplify the analysis:

$$B_r(\hat{\Theta}^k, \Theta_0) = \sum_{i=1}^{s} p_{ri}(\Theta_0)\log \frac{p_{ri}(\hat{\Theta}^k)}{p_{ri}(\Theta_0)}$$

$$= \sum_{i=1}^{s} p_{ri}(\Theta_0)\log \frac{p_{ri}(\hat{\Theta}^k)}{p_{ri}(\Theta^k)} - \sum_{i=1}^{s} p_{ri}(\Theta_0)\log \frac{p_{ri}(\Theta_0)}{p_{ri}(\Theta^k)}$$

$$= B_r(\hat{\Theta}^k, \Theta^k) - B_r(\Theta_0, \Theta^k) \tag{2.13}$$

where $\Theta^k \in R^k$. Both entropy measures are measured with respect to the density $p_{ri}(\Theta^k)$ and Θ^k is arbitrary. We will in the sequel pick Θ^k in a particular manner which clarifies and simplifies the development. The decomposition of (2.13) is done to clarify the exposition and to make clear the crucial role played by the number of parameters k. The summations in (2.13) are taken with respect to the true density on the predictive sample while $\hat{\Theta}^k$ is the estimate computed on the informative sample. Thus, $B_r(\hat{\Theta}^k, \Theta_0)$ is a measure of the information between the estimated density and the true density on the predictive sample. Since the informative sample is known but the predictive sample is not we will use statistical mean values in the sequel.

In order to evaluate $B_r(\hat{\Theta}^k, \Theta^k)$ and $B_r(\Theta_0, \Theta^k)$ we will expand around the actual probabilities on the informative sample.

Evaluation of $B_r(\hat{\Theta}^k, \Theta^k)$

From (2.13):

$$B_r(\hat{\Theta}^k, \Theta^k) = \sum_{i=1}^{s} p_{ri}(\Theta_0) \; [\log p_{ri}(\hat{\Theta}^k) - \log p_{ri}(\Theta^k)] \tag{2.14}$$

Define the sampling error between the informative and predictive probabilities as

$$e_i(\Theta) = p_{ri}(\Theta) - p_{qi}(\Theta) \tag{2.15}$$

Expanding the log term to second order yields

$$\log p_{ri}(\hat{\Theta}^k) = \log p_{qi}(\Theta^k) + \frac{\partial \log p_{qi}(\Theta^k)}{\partial p_{qi}} \; e_i(\Theta^k)$$

$$+ \frac{1}{2} \frac{\partial^2 \log p_{qi}(\Theta^k)}{\partial p_{qi}{}^2} \; e_i{}^2(\Theta^k) + \frac{\partial \log p_{qi}(\Theta^k)}{\partial \Theta^k} (\hat{\Theta}^k - \Theta^k)$$

$$+ \frac{1}{2} (\hat{\Theta}^k - \Theta^k)^T \frac{\partial^2 \log p_{qi}(\Theta^k)}{\partial \Theta^{k2}} (\hat{\Theta}^k - \Theta^k) + (\hat{\Theta}^k - \Theta^k)^T \frac{\partial^2 \log p_{qi}(\Theta^k)}{\partial \Theta^k \; \partial p_{qi}} \; e_i(\Theta^k) \tag{2.16}$$

Thus

$$B_r(\hat{\Theta}^k, \Theta^k) = \sum_{i=1}^{s} p_{ri}(\Theta_0) \frac{\partial \log p_{qi}(\Theta^k)}{\partial \Theta^k} (\hat{\Theta}^k - \Theta^k)$$

$$+ \frac{1}{2} \sum_{i=1}^{s} p_{ri}(\Theta_0) \; (\hat{\Theta}^k - \Theta^k)^T \frac{\partial^2 \log p_{qi}(\Theta^k)}{\partial \Theta^{k2}} (\hat{\Theta}^k - \Theta^k)$$

$$+ \sum_{i=1}^{s} p_{ri}(\Theta_0) \; (\hat{\Theta}^k - \Theta^k)^T \frac{\partial^2 \log p_{qi}(\Theta^k)}{\partial \Theta^k \; \partial p_{qi}} \; e_i(\Theta^k) \tag{2.17}$$

This expression can be further simplified by utilizing the fact that, since $\hat{\Theta}^k$ is a maximum-likelihood estimate on the informative sample:

$$\sum_{i=1}^{s} p_{qi}(\Theta_0) \frac{\partial \log p_{qi}(\hat{\Theta}^k)}{\partial \Theta^k} = 0 \tag{2.18}$$

Expanding this around Θ^k yields

$$\sum_{i=1}^{s} p_{qi}(\Theta_0) \left[\frac{\partial \log p_{qi}(\Theta^k)}{\partial \Theta^k} + \frac{\partial^2 \log p_{qi}(\Theta^k)}{\partial \Theta^{k2}} (\hat{\Theta}^k - \Theta^k) \right] = 0 \tag{2.19}$$

Using (2.15) and (2.19) and in (2.17) yields

$$\begin{aligned} B_r(\hat{\Theta}^k, \Theta^k) = & \sum_{i=1}^{s} e_i(\Theta_0) \frac{\partial \log p_{qi}(\Theta^k)}{\partial \Theta^k} (\hat{\Theta}^k - \Theta^k) \\ & + \frac{1}{2} \sum_{i=1}^{s} e_i(\Theta_0) (\hat{\Theta}^k - \Theta^k)^T \frac{\partial^2 \log p_{qi}(\Theta^k)}{\partial \Theta^{k2}} (\hat{\Theta}^k - \Theta^k) \\ & - \frac{1}{2} \sum_{i=1}^{s} p_{qi}(\Theta_0) (\hat{\Theta}^k - \Theta^k)^T \frac{\partial^2 \log p_{qi}(\Theta^k)}{\partial \Theta^{k2}} (\hat{\Theta}^k - \Theta^k) \\ & + \sum_{i=1}^{s} [p_{qi}(\Theta_0) + e_i(\Theta_0)] (\hat{\Theta}^k - \Theta^k)^T \frac{\partial^2 \log p_{qi}(\Theta^k)}{\partial \Theta^k \, \partial p_{qi}} e_i(\Theta_0) \end{aligned} \tag{2.20}$$

where we have assumed $e_i(\Theta^k) \approx e_i(\Theta_0)$.

The error $e_i(\Theta_0)$ is the difference of two probabilities, which are binomially distributed, by construction:

$$e_i(\Theta_0) = p_{ri}(\Theta_0) - p_{qi}(\Theta_0)$$

Furthermore $\sum_{i=1}^{s} e_i(\Theta_0) = 0$, by definition.

Since we are assuming here that $p_{ri}(\Theta_0)$ and $p_{qi}(\Theta_0)$ are independent samples from the same underlying distribution, $e_i(\Theta_0)$ is unbiased:

$$E\left\{e_i(\Theta_0)\right\} = 0 \tag{2.21}$$

where $E\{\ \}$ denotes expectation with respect to all underlying random variables. Recalling that the informative sample is of size n_q and the predictive sample is of size n_r, $p_{qi}(\Theta_0)$ has approximate variance

$$\text{var}\ (p_{qi}(\Theta_0)) = \frac{1}{n_q}\, p_i(\Theta_0)\ [1 - p_i(\Theta_0)]$$

and $p_{ri}(\Theta_0)$ has variance

$$\text{var}\ (p_{ri}(\Theta_0)) = \frac{1}{n_r}\, p_i(\Theta_0)\ [1 - p_i(\Theta_0)]$$

Thus

$$\text{var}\ (e_i(\Theta_0)) = \frac{1}{\overline{n}}\, p_i(\Theta_0)\ [1 - p_i(\Theta_0)] \tag{2.22}$$

where $\overline{n} = n_q n_r\ /\ (n_q + n_r)$. The expected value of $B_r(\hat{\Theta}^k, \Theta^k)$ can now be written in simplified form by using

$$\frac{\partial^2 \log\, p_{qi}(\Theta^k)}{\partial\Theta^k\, \partial p_{qi}} \approx -\ \frac{1}{p_i(\Theta_0)}\ \frac{\partial \log\, p_i(\Theta^k)}{\partial\Theta^k}$$

The result is that the expected value of $B_r(\hat{\Theta}^k, \Theta^k)$ is

$$E\left\{B_r(\hat{\Theta}^k, \Theta^k)\right\}$$

$$= -\frac{1}{2} E\left\{\sum_{i=1}^{s} p_i(\Theta_0)\ (\hat{\Theta}^k - \Theta^k)^T \frac{\partial^2 \log p_i(\Theta^k)}{\partial \Theta^{k2}} (\hat{\Theta}^k - \Theta^k)\right\}$$

$$-\frac{1}{\overline{n}} \sum_{i=1}^{s} E\left\{(\hat{\Theta}^k - \Theta^k)^T \frac{\partial \log p_i(\Theta^k)}{\partial \Theta^k}\right\} \tag{2.23}$$

In the sequel we will choose $\Theta^k = \Theta^{*k}$ so that Θ^{*k} is a minimum-variance estimate of Θ_0. This results in the second term being much smaller than the first term for reasonably large values of $\overline{n}/s$. We will explicitly neglect this term in the sequel.

Evaluation of $B_r(\Theta_0, \Theta^{*k})$

From (2.23)

$$B_r(\Theta_0, \Theta^k) =$$

$$-\frac{1}{2} \sum_{i=1}^{s} p_i(\Theta_0)\ (\Theta_0 - \Theta^k)^T \frac{\partial^2 \log p_i(\Theta_0)}{\partial p_i{}^2} (\Theta_0 - \Theta^k) \tag{2.24}$$

$$-\frac{1}{2} (\Theta_0 - \Theta^k)^T\ I(\Theta_0)\ (\Theta_0 - \Theta^k) \tag{2.25}$$

where $I(\Theta_0)$ is the information matrix

$$I(\Theta_0) = \sum_{i=1}^{s} p_i(\Theta_0) \frac{\partial^2 \log p_i(\Theta_0)}{\partial p_i^2} \tag{2.26}$$

In (2.23) both Θ^k and $\hat{\Theta}^k$ are k-dimensional parameter vectors. Here, however, Θ_0 is an m-dimensional vector $(m > k)$. To handle this situation, we write $\Theta_0 - \Theta^k \in R^m$ as

$$\Theta_0 - \Theta^k = \begin{bmatrix} \Theta_0{}^k - \Theta^k \\ \\ \tilde{\Theta}_0 \end{bmatrix}$$

where $\Theta_0{}^k \in R^k$, $\tilde{\Theta}_0 \in R^{m-k}$

setting

$$J(\Theta^k) = \frac{1}{2} (\Theta_0 - \Theta^k)^T I(\Theta_0) \Theta_0 - \Theta^k)$$

and minimizing with respect to Θ^k yields

$$\Theta^{*k} = \Theta_0{}^k - I_{11}{}^{-1} (\Theta_0) I_{12} (\Theta_0) \tilde{\Theta}_0 \tag{2.27}$$

where we have partitioned $I(\Theta_0)$ as

$$I(\Theta_0) = \begin{bmatrix} I_{11}(\Theta_0) & I_{12}(\Theta_0) \\ \\ I_{12}^T(\Theta_0) & I_{22}(\Theta_0) \end{bmatrix}$$

The minimum value of J is

$$J(\Theta^{*k}) = \frac{1}{2}\ \tilde{\Theta}_0^T\ [I_{22}(\Theta_0) - I_{12}(\Theta_0)^T\ I_{11}(\Theta_0)^{-1}\ I_{12}(\Theta_0)]\ \tilde{\Theta}_0$$

If we partition the covariance matrix

$$P(\Theta_0) = I(\Theta_0)^{-1}$$

$$= \begin{bmatrix} P_{11}(\Theta_0) & P_{12}(\Theta_0) \\ P_{12}^T(\Theta_0) & P_{22}(\Theta_0) \end{bmatrix}$$

then

$$J(\Theta^{*k}) = \frac{1}{2}\ \tilde{\Theta}_0^T\ \ P_{22}(\Theta_0)^{-1}\ \tilde{\Theta}_0$$

where $P_{22}(\Theta_0)^{-1} = I_{22} - I_{12}^T\ I_{11}^{-1}\ I_{12}$

Since

$$P(\Theta_0) = \sum_{i=1}^{s} P_i(\Theta)\ (\Theta_0 - \Theta^{*k})\ (\Theta_0 - \Theta^{*k})^T$$

$$\approx E\ [(\Theta_0 - \Theta^{*k})\ (\Theta_0 - \Theta^{*k})^T]$$

we get, finally,

$$E\ [J(\Theta^{*k})] \approx \frac{1}{2}\ (m - k)$$

or

$$E\ [B_r(\Theta_0,\ \Theta^{*k})] = \frac{1}{2}\ (k-m) \tag{2.28}$$

2.4 Unbiased Estimate of Entropy

From (2.23)

$$E\left\{B_r(\hat{\Theta}^k, \Theta^k)\right\} = -\frac{1}{2}(\hat{\Theta}^k - \Theta^k)^T I(\Theta^k)(\hat{\Theta}^k - \Theta^k)$$

where $I(\Theta^k)$ is the kxk information matrix

$$I(\Theta^k) = E\left\{\sum_{i=1}^{s} p_i(\Theta_0) \frac{\partial^2 \log p_i(\Theta^k)}{\partial \Theta^{k2}}\right\}$$

and $p_i(\Theta_0)$ is given in (2.6). Using (1.5) and (1.6) we see that

$$E\left\{B_r(\hat{\Theta}^k, \Theta^k)\right\} = -\frac{1}{2} \operatorname{tr} I_k$$

$$= -\frac{k}{2} \qquad (2.29)$$

where I_k is the kxk identity matrix.

Combining (2.29) and (2.28) yields

$$E\left\{B_r(\hat{\Theta}^k, \Theta_0)\right\} = \frac{m}{2} - k \qquad (2.30)$$

where $\hat{\Theta}^k$ is the maximum likelihood estimate ($\hat{\Theta}^k \varepsilon R^k$). This represents a bias in the maximized log-likelihood function, with the result that our goal is to pick k such that

$$\sum_{i=1}^{s} p_{qi}(\Theta_0) \log p_{qi}(\hat{\Theta}^k) + \frac{m}{2} - k$$

is maximized. By reference to (2.10) and (2.12), this is equivalant asymptotically to picking k such that

$$\sum_{i=1}^{n_q} \log p(x_i \mid \hat{\Theta}^k) + \frac{m}{2} - k$$

is maximized. Since m is a constant here, the equivalent goal is to minimize

$$AIC(k) = -2 \sum_{i=1}^{n_q} \log p(x_i \mid \hat{\Theta}^k) + 2k \qquad (2.31)$$

with respect to k, which is Akaike's AIC criterion.

3. CANONICAL VARIATES ANALYSIS

We now consider the linear prediction problem using the canonical variates analysis approach.

Let the past be represented as a column vector P(t) defined by

$$P(t) = \begin{bmatrix} y(t) \\ y(t-1) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}_{n \times 1}$$

and define the future as a column vector

$$F(t) = \begin{bmatrix} y(t+1) \\ y(t+2) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}_{m \times 1} \qquad m \leq n$$

where y(t) is the r-dimensional observed output at time t. Our goal is to predict the future F(t) given P(t).

We now consider the canonical variate analysis in a form that allows us to explicitly show the optimality properties of the method.

Consider nonsingular transformations of the past and future

$$\underset{n \times 1}{c(t)} = \underset{n \times n}{J} \; \underset{n \times 1}{P(t)} \tag{3.1}$$

$$\underset{m \times 1}{d(t)} = \underset{m \times m}{L} \; \underset{m \times 1}{F(t)} \tag{3.2}$$

and form a $k^{\underline{th}}$ order estimate of F(t)

$$\hat{F}_k(t) = \sum_{i=1}^{k} a_i \; c_i(t) \qquad (3.3)$$

where $\{a_i\}$ are mxl vectors and c_i is the $i^{\underline{th}}$ component of c (a scalar). Since a_i is fixed, only $c_i(t)$ depends on the data. Since J is only constrained to be nonsingular, we can use a very general form for it. Without loss of generality we can specify that

$$E\;[c(t)\;c(t)^T] = I_{nxn} \qquad (3.4)$$

Let B be an orthonormal matrix:

$$B_{nxn}^{T} \;\; B_{nxn} \; = I_{nxn} \qquad (3.5)$$

Then

$$J\; S_{pp}\; J^T = B^T\; B \qquad (3.6)$$

where $S_{pp} = E\;[P(t)\;P(t)^T]$

This has a solution

$$J = B^{T}\; S_{pp}^{-1/2} \qquad (3.7)$$

Now

$$c_i = J_i^{T}\; P(t) \qquad (3.8)$$

where J_i^{T} is the $i^{\underline{th}}$ row of J;

$$J_i^T = b_i^T S_{pp}^{-1/2} \tag{3.9}$$

and

$$B = [b_1 \ b_2 \ . \ . \ . \ b_n]_{nxn} \tag{3.10}$$

Thus

$$c_i = b_i^T S_{pp}^{-1/2} \ P(t) \tag{3.11}$$

and the estimate $\hat{F}_k(t)$ is

$$\hat{F}_k(t) = \left[\sum_{i=1}^{k} a_i \ b_i^T \right] S_{pp}^{-1/2} \quad P(t) \tag{3.12}$$

$$\triangleq Q_k \ S_{pp}^{-1/2} \ P(t)$$

where

$$Q_k = \sum_{i=1}^{k} a_i \ b_i^T \tag{3.13}$$

Note that Q_k has maximum rank k.

The prediction error is

$$e_k(t) = Q_k \ S_{pp}^{-1/2} \ P(t) - F(t) \tag{3.14}$$

We now form a quadratic cost function

$$L_k = E\,[e_k^T(t)\, W^{-1}\, e_k(t)]$$

$$= \operatorname{tr} W^{-1}\, E\,\{[Q_k\, S_{pp}^{-1/2}\, P(t) - F(t)]\, [P^T(t)\, S_{pp}^{-1/2}\, Q_k^T - F^T(t)]\}$$

$$= \operatorname{tr}(W^{-1}\, Q_k\, Q_k^T)$$

$$- 2\operatorname{tr}(W^{-1}\, Q_k\, S_{pp}^{-1/2}\, S_{pf}) + \operatorname{tr}(W^{-1}\, S_{ff}) \qquad (3.15)$$

where $S_{pf} = E\,[P(t)\, F(t)^T]$, $S_{ff} = E\,[F(t)\, F(t)^T]$

In order to handle the orthonormality constraints we add the constraint equations via Lagrange multipliers to form the augmented cost

$$\tilde{L}_k = L_k + \sum_{i=1}^{k} \lambda_i\, (b_i^T\, b_i - 1) \qquad (3.16)$$

where $\{\lambda_i\}$ are Lagrange multipliers. Thus

$$\tilde{L}_k = \operatorname{tr}\,\{W^{-1} \sum_{i=1}^{k} a_i\, b_i^T \sum_{j=1}^{n} b_j\, a_j^T\}$$

$$- 2\,\operatorname{tr}\,\{W^{-1} \sum_{i=1}^{k} a_i\, b_i^T\, S_{pp}^{-1/2}\, S_{pf}\}$$

$$+ \operatorname{tr}\,\{W^{-1}\, S_{ff}\}$$

$$+ \sum_{i=1}^{k} \lambda_i\, (b_i^T\, b_i - 1) \qquad (3.17)$$

Using $b_i^T\, b_j = \delta_{ij}$, with δ the Kroneker delta function, (3.18)

and rearranging gives

$$\tilde{L}_k = \sum_{i=1}^{k} a_i^T W^{-1} a_i$$

$$- 2 \sum_{i=1}^{k} b_i^T S_{pp}^{-1/2} S_{pf} W^{-1} a_i$$

$$+ tr(W^{-1} S_{ff}) + \sum_{i=1}^{k} \lambda_i (b_i^T b_i - 1) \tag{3.19}$$

Taking partial derivatives:

$$\frac{\partial \tilde{L}_k}{\partial a_i} = 2 a_i^T W^{-1} - 2 b_i^T S_{pp}^{-1/2} S_{pf} W^{-1} \tag{3.20}$$

$$\frac{\partial \tilde{L}_k}{\partial b_i} = -2 a_i^T W^{-1} S_{pf}^T S_{pp}^{-1/2} + 2 \lambda_i b_i^T \tag{3.21}$$

Thus, the first order necessary conditions for minimizing $\tilde{L}_k$ are

$$a_i^* = S_{pf}^T S_{pp}^{-1/2} b_i^* \tag{3.22}$$

$$\lambda_i b_i^* = S_{pp}^{-1/2} S_{pf} W^{-1} a_i^* \tag{3.23}$$

for $i = 1, 2, \ldots, k$.

Eliminating a_i^*:

$$\lambda_i b_i^* = S_{pp}^{-1/2} S_{pf} W^{-1} S_{pf}^T S_{pp}^{-1/2} b_i^* \tag{3.24}$$

which is an eigenequation.

The first term of (3.19) becomes

$$\sum_{i=1}^{k} a_i^{*T} W^{-1} a_i^{*}$$

$$= \sum_{i=1}^{k} b_i^{*T} S_{pp}^{-1/2} S_{pf} W^{-1} S_{pf}^{T} S_{pp}^{-1/2} b_i^{*}$$

$$= \sum_{i=1}^{k} \lambda_i \tag{3.25}$$

The second term of (3.19) becomes

$$-2 \sum_{i=1}^{k} b_i^{*T} S_{pp}^{-1/2} S_{pf} W^{-1} S_{pf}^{T} S_{pp}^{-1/2} b_i^{*}$$

$$= -2 \sum_{i=1}^{k} \lambda_i \tag{3.26}$$

Thus, the optimized cost is

$$L_k^{*} = \mathrm{tr}(W^{-1} S_{ff}) - \sum_{i=1}^{k} \lambda_i \tag{3.27}$$

Now let

$$R = S_{pp}^{-1/2} S_{pf} W^{-1/2} \qquad (n\mathrm{x}m) \quad n>m \tag{3.28}$$

From (3.24),

$$b_i^{*T} R R^{T} b_i^{*} = \lambda_i \tag{3.29}$$

By using a singular value decomposition on R:

$$R = U D V^T \tag{3.30}$$

$$V^T V = I, \; U^T U = I \tag{3.31}$$

$$D = \begin{bmatrix} \gamma_1 & & 0 \\ & \cdot & \\ & & \cdot \\ 0 & & \gamma_m \\ \hline & 0 & \end{bmatrix} \tag{3.32}$$

where $\gamma_1 > \gamma_2 > \ldots \gamma_m$

Then $$R R^T = U D V^T V D^T U^T$$

$$= U D D^T U^T \tag{3.33}$$

Then, from (3.29)

$$b_i^{*T} U D D^T U^T b_i^* = \lambda_i \tag{3.34}$$

Thus b_i^* is the eigenvector of $U D D^T U^T$ whose eigenvalue is λ_i.

Now let

$$U = [U_1 \; U_2 \; \ldots \; U_n] \tag{3.35}$$

where the U_i are mutually orthogonal unit vectors by construction. But the matrix $U D D^T U^T$ has eigenvectors U_i and associated eigenvectors γ_i^2 since

$$U_i^T U D D^T U^T U_j = \gamma_i^2 \delta_{ij} \tag{3.36}$$

Thus

$$\gamma_i^2 = \lambda_i, \; U_i = b_i^* \tag{3.37}$$

and

$$a_i^* = S_{pf}^T \; S_{pp}^{-1/2} \; U_i \tag{3.38}$$

By using (3.37) in (3.27) we get

$$L_k^* = tr(W^{-1} \; S_{ff}) - \sum_{i=1}^{k} \gamma_i^2 \tag{3.39}$$

and we see that the cost is minimized by using the k largest canonical variances, $\gamma_1^2 > \gamma_2^2 > \gamma_3^2 > . . > \gamma_k^2$.

We can now write the optimal forecast as

$$F_k^*(t) = \sum_{i=1}^{k} a_i^* \; c_i^*(t)$$

$$= \sum_{i=1}^{k} S_{pf}^T \; S_{pp}^{-1/2} \; U_i \, U_i^T \; S_{pp}^{-1/2} \; P(t)$$

$$= S_{pf}^T \, S_{pp}^{-1/2} \; (\sum_{i=1}^{k} U_i \; U_i^T) \; S_{pp}^{-1/2} \; P(t) \tag{3.40}$$

Thus, if we denote the optimal weighting matrix by A_k^*:

$$F_k^*(t) = A_k^* \; P(t) \tag{3.41}$$

$$A_k^* = S_{pf}^T \; S_{pp}^{-1/2} \left[\sum_{i=1}^{k} U_i \; U_i^T \right] S_{pp}^{-1/2} \tag{3.42}$$

Note that

$$A_n^* = S_{pf}^T S_{pp}^{-1} \tag{3.43}$$

To determine L (cf (3.2)), we can use the condition

$$E(cd^T) = D \tag{3.44}$$

or

$$J^* S_{pf} L^T = D \tag{3.45}$$

From (3.7),

$$U^T S_{pp}^{-1/2} S_{pf} L^T = D \tag{3.46}$$

But

$$D = U^T R V$$

$$= U^T S_{pp}^{-1/2} S_{pf} W^{-1/2} V \tag{3.47}$$

Comparing (3.46) and (3.47) gives

$$L^T = W^{-1/2} V \text{ , or}$$

$$L = V^T W^{-1/2} \tag{3.48}$$

Note that A_k^*, the optimal gain matrix is of dimension mxn but has a maximum rank of k.

Note that $k \leq m$ since the symmetric matrix in the eigenequation (3.24) has rank $\leq m$. This is very important, as it implies that we need to make the dimension of the future vector (m) at least as large as the maximum expected order of the estimator.

An efficient computation of A_k^* is

$$d_i^* = S_{pp}^{-1/2} U_i \; ; \; S_{pp}^{-1/2} \quad \text{symmetric}$$

$$a_i^* = S_{pf}^T d_i^*$$

$$A_k^* = \sum_{i=1}^{k} a_i^* d_i^{*T}$$

Cholesky Form

The cholesky factorization of a positive-definite matrix is an attractive way of computing a square root matrix. Let

$$S_{pp}^{-1} = Q Q^T$$

Then we get the following relations

$$a_i^* = S_{pf}^T Q U_i$$

$$\lambda_i^* U_i = Q^T S_{pf} W^{-1} S_{pf}^T Q U_i$$

$$R = Q^T S_{pf} W^{-1/2}$$

$$F_k^*(t) = S_{pf}^T Q \left(\sum_{i=1}^{k} U_i U_i^T \right) Q^T P(t)$$

$$A_k^* = S_{pf}^T Q \left(\sum_{i=1}^{k} U_i U_i^T \right) Q^T$$

Trucated Predictor

In the sequel, we will be restricting the total number of parameters allowed in the predictor. The question arises as how to best truncate the prediction equations. Our approach is to use only the most recent past values. For example, suppose we have used m = 5 in our analysis, but wish only to use a one-step-ahead predictor with k parameters. Then our predictor uses only the first k elements of the first row of A_k^*.

Inclusion of Known Inputs

If we have an unknown system with measured outputs y(t) and measured inputs u(t), the analysis of this section holds with only slight modifications. If we augment the past vector as

$$P(t) = \begin{bmatrix} y(t) \\ u(t) \\ y(t-1) \\ u(t-1) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \qquad (3.49)$$

then all of the analyes of this section holds and the predicted values of y(t) depend on both past values of y(t) and on past values u(t).

4. DIRECT DETERMINATION OF STATE-SPACE MATRICES

We now consider the problem of determining the state-space matrices directly from the linear prediction solution. Recall from Section 3:

$$P(t) = \begin{bmatrix} y(t) \\ y(t-1) \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}_{n\times 1} \tag{4.1}$$

$$\hat{F}(t) = A\, P(t) \tag{4.2}$$

If we restrict our problem to a one-step-ahead prediction of y(t), then

$$\hat{y}(t+1 \mid t) = A\, P(t) \tag{4.3}$$

where A is mxn.

We can write a recursion for P(t) as follows:

$$P(t+1) = M\, P(t) + T\, y(t+1) \tag{4.4}$$

where

$$M = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ I & 0 & \cdots & \cdots & 0 \\ 0 & I & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & 0 & I & 0 \end{bmatrix} \tag{4.5}$$

where all submatrices are mxm.

$$T = \begin{bmatrix} I_{mxm} \\ 0_{(n-m)xm} \end{bmatrix} \tag{4.6}$$

so that P(t) is in recursive form with a driving term y(t+1).

The state-space formulation employs a Kalman filter for updating. The equations for the time-invariant filter at steady state are

$$\hat{y}(t+1 \mid t) = H\ \hat{x}(t+1 \mid t) \tag{4.7}$$

$$\hat{x}(t+1 \mid t) = \Phi\ \hat{x}(t \mid t) \tag{4.8}$$

$$\hat{x}(t+1 \mid t+1) = \hat{x}(t+1 \mid t) + K\ [y(t+1) - \hat{y}(t+1 \mid t)] \tag{4.9}$$

Combining (4.7) - (4.9) yields

$$\begin{cases} \hat{y}(t+1 \mid t) = H\ \Phi\ \hat{x}(t \mid t) & (4.10) \\ \hat{x}(t+1 \mid t+1) = (I-KH)\ \Phi\ \hat{x}(t \mid t) + K\ y(t+1) & (4.11) \end{cases}$$

as the state-space equation set. The linear prediction set is

$$\begin{cases} \hat{y}(t+1 \mid t) = A\ P(t) & (4.12) \\ P(t+1) = M\ P(t) + T\ y(t+1) & (4.13) \end{cases}$$

What we seek to do is <u>match</u> these two pairs of equations by finding the state-space matrices H, Φ, K which give the best "fit" to the linear prediction equations.

Equations (4.7) - (4.9) can also be put in the form

$$\hat{y}(t+1 \mid t) = H\,\hat{x}(t+1 \mid t) \tag{4.14}$$

$$\hat{x}(t+1 \mid t) = \Phi(I - KH)\hat{x}(t \mid t-1) + \Phi\,K\,y(t) \tag{4.15}$$

We can solve the problem operationally by using solutions involving delay operators.

We will first solve the linear prediction equations. From (4.13),

$$P(t+1) = zM\,P(t+1) + T\,y(t+1) \tag{4.16}$$

where z is the delay operator: $z\,P(t+1) = P(t)$

Solving (4.16):

$$P(t+1) = (I-zM)^{-1}\,T\,y(t+1) \tag{4.17}$$

Thus, from (4.12):

$$\hat{y}(t+1 \mid t) = A\,(I-zM)^{-1}\,T\,y(t) \tag{4.18}$$

We can also solve the state-space equations in the same way.

From (4.10) and (4.11) we get

$$\hat{x}(t \mid t) = [I - (I-KH)\,\Phi z]^{-1}\,K\,y(t) \tag{4.19}$$

so that

$$\hat{y}(t+1 \mid t) = H\,\Phi\,[I - (I-KH)\,\Phi z]^{-1}\,K\,y(t) \tag{4.20}$$

while from (4.14) and (4.15) we get

$$\hat{x}(t+2 \mid t+1) = [I - \Phi(I-KH)z]^{-1} \Phi K y(t+1) \tag{4.21}$$

$$\hat{y}(t+1 \mid t) = H [I - \Phi(I-KH)z]^{-1} \Phi K y(t) \tag{4.22}$$

If we could get a perfect match between the linear prediction and the state-space predictors, then the following equation would be satisfied

$$A (I-Mz)^{-1} T = H \Phi[I-(I-KH)\Phi z]^{-1} K \tag{4.23}$$

or, equivalently

$$A (I-Mz)^{-1} T = H[I-\Phi(I-KH)z]^{-1} \Phi K \tag{4.24}$$

Using (4.23), we see that exact matching occurs, for dim (x) = n, if

$$\begin{cases} A = H \Phi & (4.25) \\ M = (I-KH) \Phi & (4.26) \\ T = K & (4.27) \end{cases}$$

Equation (4.26) can be written as

$$M = \Phi - TA, \tag{4.28}$$

so that

$$\Phi = M + TA \tag{4.29}$$

In addition, (4.25) gives

$$H = A \Phi^{-1} \tag{4.30}$$

Since H must satisfy A = H (M + TA), it is easily shown that H is in canonical form

$$H = [I_{m\times m} \; 0_{m\times(n-m)}]$$

If Φ is a valid transition matrix, it is guaranteed to be invertible. Thus we only need to guarantee the invertibility of M + TA. Using the definition of M and partitioning T and A appropriately, we see

$$\Phi_{n\times n} = \begin{bmatrix} 0_{m\times(n-m)} & 0_{m\times m} \\ I_{(n-m)\times(n-m)} & 0_{(n-m)\times m} \end{bmatrix}$$

$$+ \begin{bmatrix} I_{m\times m} \\ 0_{(n-m)\times m} \end{bmatrix} \begin{bmatrix} A_{1\,m\times(n-m)} & A_{2\,m\times m} \end{bmatrix}$$

$$= \begin{bmatrix} A_1 & A_2 \\ I & 0 \end{bmatrix} \tag{4.31}$$

The inverse is

$$\Phi^{-1} = \begin{bmatrix} 0 & I \\ A_2^{-1} & -A_2^{-1} A_1 \end{bmatrix} \tag{4.32}$$

Thus Φ^{-1} exists if A_2^{-1} exists. To check this, write equation (4.15) in partitioned form as

$$[A_1 \quad A_2] =$$

$$\begin{bmatrix} S_{pf1}^T & S_{pf2}^T{}_{p\times p} \end{bmatrix} \begin{bmatrix} W_{11} & W_{12} \\ W_{12}^T & W_{22} \end{bmatrix} \tag{4.33}$$

where

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{12}^T & W_{22} \end{bmatrix} = S_{pp}^{-1} \tag{4.34}$$

Then

$$A_2 = S_{pf1}^T W_{12} + S_{pf2}^T W_{22} \tag{4.35}$$

Now partition S_{pp} as

$$S_{pp} = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix} \tag{4.36}$$

Then

$$W_{12} = - S_{11}^{-1} S_{12} (S_{22} - S_{12}^T S_{11}^{-1} S_{12})^{-1} \tag{4.37}$$

$$W_{22} = (S_{22} - S_{12}^T S_{11}^{-1} S_{12})^{-1} \tag{4.38}$$

so that

$$A_2 = (S_{pf2}^T - S_{pf1}^T S_{11}^{-1} S_{12}) (S_{22} - S_{12}^T S_{11}^{-1} S_{12})^{-1} \tag{4.39}$$

Therefore

$$A_2^{-1} = (S_{22} - S_{12}^T S_{11}^{-1} S_{12}) (S_{pf2}^T - S_{pf1}^T S_{11}^{-1} S_{12})^{-1} \tag{4.40}$$

Solving for A_1 yields

$$A_1 = S_{pf1}^T W_{11} + S_{pf2}^T W_{12}^T \tag{4.41}$$

Using

$$W_{11} = S_{11}^{-1} + S_{11}^{-1} S_{12} (S_{22} - S_{12}^{T} S_{11}^{-1} S_{12})^{-1} S_{12}^{T} S_{11}^{-1} \quad (4.42)$$

we get

$$A_1 = S_{pf1}^{T} S_{11}^{-1} +$$

$$[S_{pf1}^{T} S_{11}^{-1} S_{12} - S_{pf2}^{T}] (S_{22} - S_{12}^{T} S_{11}^{-1} S_{12})^{-1} S_{12}^{T} S_{11}^{-1} \quad (4.43)$$

In summary, we can use the direct solution if the matrix

$$S_{ce} = S_{pf2}^{T} - S_{pf1}^{T} S_{11}^{-1} S_{12} \quad (4.44)$$

is invertible.

This exact solution is restricted to the case dim (x) = n. This implies that

$$\dim(x(t)) = \dim(P(t))$$

Truncated Filter

Given the matrices for the full-order Kalman filter Φ_{nxn}, H_{mxn}, K_{nxm}, the question arises as to whether there is a suitable truncation to a lower-oder form required to meet restrictions on the total number of parameters. Using the forms of (4.27), (4.29) and (4.30) and assuming an order $k < n$, and prediction of the first p future values ($p < m$), we truncate as follows:

(1) Φ_{kxk} is the upper left kxk submatrix of Φ_{nxn}

(2) H_{pxk} is the upper left pxk submatrix of H_{mxn}

(3) K_{kxp} is the upper left kxp submatrix of K_{nxm}

Then the Kalman filter using Φ_{kxk}, H_{pxk}, K_{kxp} yields exactly the same predictions as the truncated linear predictor of Section 3.

5. MODEL SELECTION FOR LINEAR PREDICTION

We can now consider the problem of model order selection for linear prediction problems using canonical variates analysis. Assume that the data y(t) are Gaussian and consider data on the interval [t, s] t<s with unit time increment and let d = s - t+1. The associated data over the interval are

$$Y(t,s) = \{y(t),\ y(t+1),\ .\ .\ .\ .\ y(t+d-1)\} \tag{5.1}$$

The one-step-ahead prediction error

$$e_k(t) = y(t+1) - A_k\ p(t) \tag{5.2}$$

is also Gaussian and, from (3.42) has zero mean and variance

$$E\ \{e_k(t)\ e_k(t)^T\}$$

$$= S_k(t+1 \mid t)$$

$$= S_{ff} - 2\ S_{pf}^T\ S_{pp}^{-1/2}\ U^k\ S_{pp}^{-1/2}\ S_{pf}$$

$$+ S_{pf}^T\ S_{pp}^{-1/2}\ (U^k)^2\ S_{pp}^{-1/2}\ S_{pf} \tag{5.3}$$

where

$$U^k = \sum_{i=1}^{k} U_i\ U_i^T \tag{5.4}$$

Since $(U^k)^2 = U^k$ (5.3) reduces to

$$S_k(t+1 \mid t) = S_{ff} - S_{pf}^T S_{pp}^{-1/2} U^k S_{pp}^{-1/2} S_{pf} \tag{5.5}$$

Since $S_k(t+1 \mid t)$ is constant for all t due to assumed stationarity of y(t), we can set $S_k = S_k(t+1 \mid t)$.

From (3.3) the number of parameters is mk. Thus

$$AIC(k) = \sum_{t=1}^{d} \left[\log \left| S_k \right| + e_k(t)^T S_k^{-1} e_k(t) \right] + 2\ mk$$

$$\approx d[\log \left| S_k \right| + tr\ I_m] + 2\ mk \tag{5.6}$$

where $|\cdot|$ denotes determinant and I_m is the mxm identity matrix. The AIC is thus

$$AIC(k) = d \log \left| S_k \right| + m + 2\ mk \tag{5.7}$$

Since d and m are fixed, we see that the optimum order k is the one which minimizes

$$\log \left| S_k \right| + \frac{2mk}{d} \tag{5.8}$$

for $d >> 2(m)k$ this is approximately equivalent to minimizing

$$\left| S_k \right| \left[1 + \frac{2\ mk}{d} \right] \tag{5.9}$$

which can be recognized as the forward prediction error (FPE), see Akaike (1973). In the sequel we will use (5.8) instead of (5.9) since the data interval will sometimes be relatively short.

6. DETECTION OF ABRUPT MODEL CHANGES

In this section we apply the tools developed so far to the problem of abrupt change detection. We then present some experimental results.

6.1 Algorithm Development

Consider the situation depicted in Figure 6.1, in which the true time-series model changes from M_0 to M_1 at time $t-d_1$, where t is the present time. Suppose

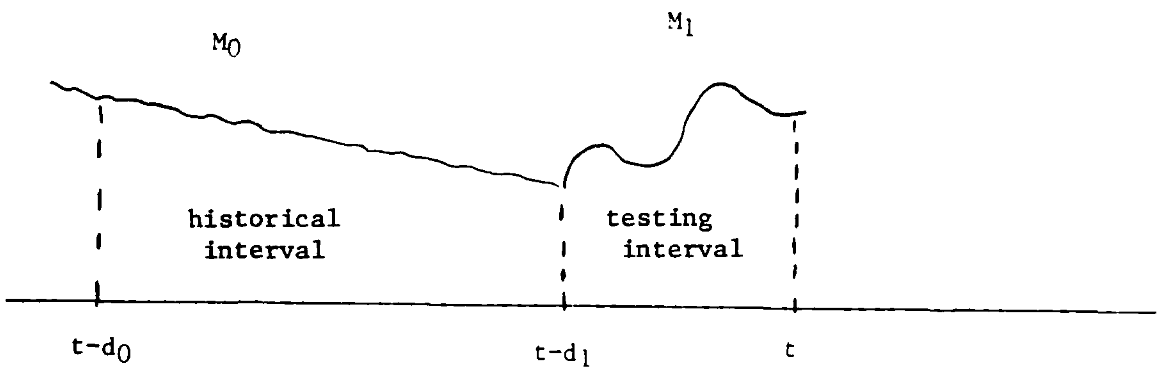


Figure 6.1 Changing Time Series Model

that we have data back to time $t-d_0$ and that the true model is M_0 in the interval $(t-d_0, t-d_1)$.

We wish to detect this change in the model. In this example, fitting a single model to data over the interval $(t-d_0, t)$ should result in greater fit errors than fitting one model over the interval $(t-d_0, t-d_1)$ and another model over the interval $(t-d_1, t)$. The crucial issue is to determine an appropriate selection measure so as to be sensitive to

changing models, while at the same time not being too sensitive to noise. Over-sensitivity to noise will result in deciding that model changes have occurred when, in fact, they have not. Low sensitivity to model changes will result in missing changes which have occurred in the model. Another obvious problem is how best to select the testing intervals, $(t-d_0, t)$ and $(t-d_1, t)$, to minimize the time required to achieve accurate detection. We consider first the selection measure.

Over the interval $(t-d_0, t)$ we can find the model which minimizes the AIC:

$$AIC(k) = -2 \sum_{i=1}^{d_0} \log p(e(t-d_0+i) \mid \hat{\Theta}^k) + 2M(k) \tag{6.1}$$

where M(k) is the number of independently adjustable parameters and where we have assumed a sampling time increment of one, for convenience.

If we now divide the interval $(t-d_0, t)$ into two subintervals, $(t-d_0, t-d_1)$ and $(t-d_1, t)$, we determine minimum AIC models for each subinterval

$$AIC_0(k) = -2 \sum_{i=1}^{d_0-d_1} \log p(e(t-d_0+i) \mid \hat{\Theta}^k) + 2M(k) \tag{6.2}$$

$$AIC_1(k) = -2 \sum_{i=d_0-d_1+1}^{d_0} \log p(e(t-d_0+i) \mid \hat{\Theta}^k) + 2M(k) \tag{6.3}$$

Now assume that

$$k^* = \arg\min AIC(k)$$

$$k_0^* = \arg\min AIC_0(k)$$

$$k_1^* = \arg\min AIC_1(k)$$

and let the corresponding models be parametrized by $\hat{\Theta}^{k^*}$, $\hat{\Theta}_0^{k_0^*}$, $\hat{\Theta}_1^{k_1^*}$ respectively.

Then the model selection criterion is based on comparing $AIC(k^*)$ with $AIC_0(k_0^*) + AIC(k_1^*)$ and selecting the model(s) which give the least value. We can simplify the calculation in the case that $d_0 >> d_1$ and the model does not change too much, in which case we expect that $k^* \approx k_0^*$, $\Theta^{k^*} \approx \Theta_0^{k_0^*}$. In this case we can define the AIC difference as

$$\Delta AIC^* = AIC(k^*) - AIC_0(k_0^*) - AIC_1(k_1^*)$$

$$= -2 \sum_{i=d_0-d_1+1}^{d_0} \log p(e(t-d_0+i) \mid \hat{\Theta}^{k^*}$$

$$+ 2 \sum_{i=d_0-d_1+1}^{d_0} \log p(e(t-d_0+i) \mid \Theta_1^{k_1^*}) - 2\, M(k_1^*) \qquad (6.4)$$

and the decision rule is

$$\Delta AIC^* \begin{cases} < 0 \text{ ; declare "no change"} \\ > 0 \text{ ; declare "change"} \end{cases} \qquad (6.5)$$

Note that ΔAIC^* may be written as

$$\Delta AIC^* = 2 \sum_{i=1d_0-d_1+1}^{d_0} \log \frac{p(e(t-d_0+i) \mid \hat{\Theta}_1^{k_1^*})}{p(e(t-d_0+i) \mid \hat{\Theta}^{k^*})} - 2\, M(k_1^*) \qquad (6.6)$$

which is the likelihood ratio in favor of the best model in the interval $(t-d_1, t)$ to the best historical model evaluated over the same interval, but biased off by the number of parameters of the best model in the interval $(t-d_1, t)$.

If we specialize this result to the linear prediction problem under study here, we see that

$$\Delta AIC^* \approx d_1 \left\{\log \left| S(k^*) \right| + tr\ \tilde{S}(k^*)\ S(k^*)^{-1}\right\}$$
$$- d_1 \left\{\log \left| S_1(k_1{}^*) \right| + m\right\}$$
$$- 2\ mk_1{}^* \tag{6.7}$$

where $S(k^*)$ is the theoretical covariance matrix of prediction errors for the historical model fitted on the interval $(t-d_0, t-d_1)$, $S_1(k_1{}^*)$ is the theoretical covariance matrix of prediction errors for the model fitted to the data on the interval $(t-d_1, t)$, and $\tilde{S}(k^*)$ is the actual covariance matrix of prediction errors for the historical model, evaluated on the interval $(t-d_1, t)$. Now let $\Delta\tilde{S}(k^*) = \tilde{S}(k^*) - S(k^*)$.

Then

$$\Delta AIC^* = d_1\ \log \left\{ \frac{\left| S(k^*) \right|}{\left| S_1(k_1{}^*) \right|} \exp \left[\frac{-\ 2\ \ mk_1{}^*}{d_1} + tr\ \Delta\tilde{S}(k^*) S(k^*)^{-1} \right] \right\} \tag{6.8}$$

Thus our decision parameter is

$$\gamma = \log \frac{|S(k^*)|}{|S_1(k_1^*)|} - \frac{2\ mk_1^*}{d_1} + \mathrm{tr}\ \Delta\tilde{S}(k^*)S(k^*)^{-1} \tag{6.9}$$

and the decision rule is

$$\gamma \begin{cases} < 0 \ ; \text{ declare "no change"} \\ > 0 \ ; \text{ declare "change"} \end{cases} \tag{6.10}$$

6.2 Experimental Results

The abrupt change detector was tried on a changing autoregressive model. On the interval $t \in [1, d_0]$, the actual model was

$$y(t) = 1.65\ y(t-1) - 0.665\ y(t-2) + u(t) \qquad \text{(Model 1)} \tag{6.11}$$

where u(t) was zero-mean white Gaussian noise with variance of 1. This model has two real stable poles at 0.95 and 0.7. The actual model was then changed to

$$y(t) = 2.5\ y(t-1) - 2.11\ y(t-2) + 0.595\ y(t-3) + u(t) \ \text{(Model 2)} \tag{6.12}$$

on the interval $t \in [d_0 + 1, d_0 + d,]$. This model has three poles at 0.7, 0.9 + 0.2i, 0.9 - 0.2i.

The first trial used $d_0 = 80$, $d_1 = 20$. The resulting covariance matrices on the interval $[1, d_0]$ were

$$S_{pp1} = \begin{Bmatrix} 11.0808 & 10.9642 & 10.7662 & 10.5378 & 10.2860 \\ 10.9642 & 11.0018 & 10.8992 & 10.7144 & 10.4733 \\ 10.7662 & 10.8992 & 10.9484 & 10.8566 & 10.6614 \\ 10.5378 & 10.7144 & 10.8566 & 10.9144 & 10.8144 \\ 10.2860 & 10.4733 & 10.6614 & 10.8144 & 10.8619 \end{Bmatrix}$$

$$S_{pf1} = \begin{Bmatrix} 11.0439 & 10.9089 & 10.7124 & 10.4822 & 10.2222 \\ 10.8318 & 10.6534 & 10.4500 & 10.2258 & 9.9723 \\ 10.5900 & 10.4015 & 10.1993 & 9.9753 & 9.7098 \\ 10.3510 & 10.1607 & 9.9542 & 9.7122 & 9.4267 \\ 10.0978 & 9.9061 & 9.6859 & 9.4296 & 9.1241 \end{Bmatrix}$$

$$S_{ff1} = \begin{Bmatrix} 11.1612 & 11.1214 & 10.9673 & 10.7430 & 10.4764 \\ 11.1214 & 11.2356 & 11.1760 & 10.9937 & 10.7336 \\ 10.9673 & 11.1760 & 11.2697 & 11.1831 & 10.9711 \\ 10.7430 & 10.9937 & 11.1831 & 11.2537 & 11.1474 \\ 10.4764 & 10.7336 & 10.9711 & 11.1474 & 11.2135 \end{Bmatrix}$$

By performing an SVD, we obtain

$$U_1 = \begin{Bmatrix} 0.6655 & 0.4685 & -0.4465 & 0.3351 & -0.1607 \\ 0.4239 & -0.2456 & 0.7166 & 0.4911 & 0.0726 \\ 0.3888 & 0.2135 & 0.3803 & -0.7320 & -0.3504 \\ 0.3596 & -0.1033 & -0.1198 & -0.3149 & 0.8640 \\ 0.3112 & -0.8148 & -0.3579 & -0.1072 & -0.3157 \end{Bmatrix}$$

$$D_1 = \begin{Bmatrix} 7.2182 & 0 & 0 & 0 & 0 \\ 0 & 0.1863 & 0 & 0 & 0 \\ 0 & 0 & 0.1032 & 0 & 0 \\ 0 & 0 & 0 & 0.0207 & 0 \\ 0 & 0 & 0 & 0 & 0.0165 \end{Bmatrix}$$

$$V_1 = \begin{Bmatrix} 0.4605 & -0.6945 & 0.5094 & -0.1666 & 0.1355 \\ 0.4569 & -0.2407 & -0.4337 & 0.4939 & -0.5489 \\ 0.4493 & 0.0706 & -0.5097 & -0.0174 & 0.7301 \\ 0.4398 & 0.3337 & -0.0833 & -0.7382 & -0.3787 \\ 0.4289 & 0.5860 & 0.5344 & 0.4279 & 0.0627 \end{Bmatrix}$$

The resulting values of AIC(k) for different orders k are, neglecting constants,

$$\begin{aligned} AIC(1) &= -\ 2.150 \\ AIC(2) &= -\ 2.264 \\ AIC(3) &= -\ 2.243 \\ AIC(4) &= -\ 2.195 \end{aligned}$$

so that $k^* = 2$, which is the correct order, is selected. The estimated model is $y(t) = 1.843\ y(t-1) - 1.0081\ y(t-2)$.

On the interval $[d_0 + 1, d_0 + d_1]$, the covariance matrices were

$$S_{pp2} = \begin{Bmatrix} 1.7647 & 1.7624 & 1.6008 & 1.3214 & 1.1104 \\ 1.7624 & 1.9451 & 1.9171 & 1.7039 & 1.4781 \\ 1.6008 & 1.9171 & 2.0772 & 2.0067 & 1.8375 \\ 1.3214 & 1.7039 & 2.0067 & 2.1332 & 2.0981 \\ 1.3214 & 1.7039 & 2.0067 & 2.1332 & 2.0981 \end{Bmatrix}$$

$$S_{pf2} = \begin{Bmatrix} 1.6394 & 1.4902 & 1.4677 & 1.6660 & 2.1622 \\ 1.4905 & 1.2481 & 1.1833 & 1.3743 & 1.8518 \\ 1.2626 & 1.0238 & 1.0057 & 1.2397 & 1.7310 \\ 0.9897 & 0.8111 & 0.8554 & 1.1288 & 1.6058 \\ 0.8331 & 0.7084 & 0.7783 & 1.0178 & 1.4219 \end{Bmatrix}$$

$$S_{ff2} = \begin{bmatrix} 1.7259 & 1.7733 & 1.9101 & 2.2212 & 2.7981 \\ 1.7733 & 2.1250 & 2.5776 & 3.1739 & 3.9927 \\ 1.9101 & 2.5776 & 3.4770 & 4.5339 & 5.7802 \\ 2.2212 & 3.1739 & 4.5339 & 6.1817 & 8.0268 \\ 2.7981 & 3.9927 & 5.7802 & 8.0268 & 10.5912 \end{bmatrix}$$

Performing the SVD yields.

$$U_2 = \begin{bmatrix} 0.9394 & -0.1382 & -0.2689 & 0.0560 & -0.1513 \\ 0.0073 & -0.8145 & 0.4181 & 0.3600 & 0.1791 \\ 0.1124 & -0.1507 & 0.5108 & -0.7606 & -0.3538 \\ 0.2936 & 0.5410 & 0.6957 & 0.3074 & 0.2065 \\ 0.1361 & -0.0455 & -0.0891 & -0.4408 & 0.8816 \end{bmatrix}$$

$$D_2 = \begin{bmatrix} 3.5041 & 0 & 0 & 0 & 0 \\ 0 & 0.6770 & 0 & 0 & 0 \\ 0 & 0 & 0.2473 & 0 & 0 \\ 0 & 0 & 0 & 0.0326 & 0 \\ 0 & 0 & 0 & 0 & 0.0094 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.3352 & -0.7554 & 0.4483 & -0.2558 & 0.2250 \\ 0.3611 & -0.3926 & -0.3766 & 0.6815 & -0.3305 \\ 0.4033 & -0.0185 & -0.6095 & -0.6546 & -0.1926 \\ 0.4766 & 0.2819 & -0.1621 & 0.2039 & 0.7909 \\ 0.6062 & 0.4422 & 0.5093 & 0.0107 & -0.4213 \end{bmatrix}$$

The resulting values of AIC(k) are, again neglecting constants,

$$
\begin{aligned}
AIC(1) &= -\ 0.960 \\
AIC(2) &= -\ 2.266 \\
AIC(3) &= -\ 2.323 \\
AIC(4) &= -\ 2.223 \\
AIC(5) &= -\ 2.123
\end{aligned}
$$

so that $k^* = 3$, as desired. The estimated model is

$$y(t) = 1.9456\ y(t-1) - 1.1485\ y(t-2) + 0.0483\ y(t-3)$$

Note that, with the sparse amount of data available, the coefficient errors are relatively large and the two estimated models are relatively close to each other.

The ΔAIC criterion was used to test for a change in the time series coefficients. Since we have only one output, the criterion is

$$\Delta AIC = \log \frac{S(k^*)}{S_1(k_1^*)} + \frac{\tilde{S}(k^*) - S(k^*)}{S(k^*)} - \frac{2\ k_1^*}{d1} \tag{6.13}$$

Using $k^* = 2$, $k_1^* = 3$, $S(k^*) = .0940$, $S_1(k_1^*) = .0726$, $S(k^*) = .0903$, $d_1 = 20$ yields,

$$\Delta AIC = 0.2583 - 0.0394 - 0.3 = -\ 0.0811$$

so that a "no change" decision is made, but just barely. Note that the actual covariance on the second interval using Model #1 is actually less than for the first interval, as a result of using only a small testing interval.

We next tried the test over larger intervals, keeping a 4:1 ratio between the historical interval and the testing interval. The intervals used were 160 for the historical interval and 40 for the testing interval.

The covariance matrices for the historical interval were

$$S_{pp1} = \begin{bmatrix} 6.4177 & 6.3504 & 6.2021 & 6.0182 & 5.8335 \\ 6.3504 & 6.4177 & 6.3504 & 6.2022 & 6.0183 \\ 6.2021 & 6.3504 & 6.4170 & 6.3494 & 6.2006 \\ 6.0182 & 6.2022 & 6.3494 & 6.4153 & 6.3470 \\ 5.8335 & 6.0183 & 6.2006 & 6.3470 & 6.4119 \end{bmatrix}$$

$$S_{pf1} = \begin{bmatrix} 6.3504 & 6.2014 & 6.0149 & 5.8244 & 5.6495 \\ 6.2013 & 6.0147 & 5.8242 & 5.6493 & 5.4869 \\ 6.0169 & 5.8282 & 5.6544 & 5.4919 & 5.3310 \\ 5.8316 & 5.6606 & 5.4999 & 5.3387 & 5.1657 \\ 5.6655 & 5.5089 & 5.3502 & 5.1770 & 4.9871 \end{bmatrix}$$

$$S_{ff1} = \begin{bmatrix} 6.4187 & 6.3529 & 6.2057 & 6.0202 & 5.8294 \\ 6.3529 & 6.4251 & 6.3642 & 6.2194 & 6.0332 \\ 6.2057 & 6.3642 & 6.4436 & 6.3856 & 6.2387 \\ 6.0202 & 6.2194 & 6.3856 & 6.4671 & 6.4058 \\ 5.8294 & 6.0332 & 6.2387 & 6.4058 & 6.4835 \end{bmatrix}$$

The results of the SVD were

$$U_1 = \begin{bmatrix} 0.6995 & 0.4349 & -0.2706 & 0.4797 & 0.1353 \\ 0.3997 & -0.7769 & 0.1851 & 0.3162 & -0.3202 \\ 0.3573 & -0.0762 & 0.5794 & -0.3109 & 0.6589 \\ 0.3481 & 0.3521 & 0.3443 & -0.4460 & -0.6613 \\ 0.3197 & -0.2785 & -0.6620 & -0.6118 & 0.0879 \end{bmatrix}$$

$$D = \begin{bmatrix} 5.3574 & 0 & 0 & 0 & 0 \\ 0 & 0.1611 & 0 & 0 & 0 \\ 0 & 0 & 0.1026 & 0 & 0 \\ 0 & 0 & 0 & 0.0233 & 0 \\ 0 & 0 & 0 & 0 & 0.0070 \end{bmatrix}$$

$$V = \begin{Bmatrix} 0.4693 & -0.8139 & 0.1915 & -0.2626 & 0.1082 \\ 0.4616 & -0.1011 & -0.5037 & 0.6123 & -0.3847 \\ 0.4489 & 0.3182 & -0.4470 & -0.2359 & 0.6647 \\ 0.4342 & 0.3743 & 0.1072 & -0.5518 & -0.5962 \\ 0.4204 & 0.2933 & 0.7059 & 0.4426 & 0.2075 \end{Bmatrix}$$

The values of AIC (k) were

$$\begin{aligned} AIC(1) &= -\ 2.3072 \\ AIC(2) &= -\ 2.4871 \longleftarrow \\ AIC(3) &= -\ 2.4795 \\ AIC(4) &= -\ 2.467 \\ AIC(5) &= -\ 2.4545 \end{aligned}$$

so that $k^* = 2$ is selected. The estimated model is

$$y(t) = 1.6777\ y(t-1) - 0.7178\ y(t-2)$$

which is much closer to the actual model (6.11), due to the increased data length.

The model over the testing interval was next found. The covariance matrices were

$$S_{pp2} = \begin{Bmatrix} 20.6940 & 20.2655 & 19.2378 & 17.6822 & 15.7286 \\ 20.2655 & 20.4909 & 20.0699 & 19.0219 & 17.4532 \\ 19.2378 & 20.0699 & 20.3055 & 19.8666 & 18.8081 \\ 17.6822 & 19.0219 & 19.8666 & 20.0831 & 19.6334 \\ 15.7286 & 17.4532 & 18.8081 & 19.6334 & 19.8397 \end{Bmatrix}$$

$$S_{pf2} = \begin{Bmatrix} 20.4906 & 19.6758 & 18.3328 & 16.5805 & 14.5504 \\ 19.4580 & 18.1453 & 16.4322 & 14.4431 & 12.2933 \\ 17.9266 & 16.2349 & 14.2789 & 12.1690 & 9.9935 \\ 15.9897 & 14.0540 & 11.9786 & 9.8473 & 7.7296 \\ 13.7888 & 11.7306 & 9.6332 & 7.5625 & 5.5657 \end{Bmatrix}$$

$$S_{ff2} = \begin{bmatrix} 20.9398 & 20.7220 & 19.8651 & 18.4736 & 16.6762 \\ 20.7220 & 21.1467 & 20.8735 & 19.9608 & 18.5264 \\ 19.8651 & 20.8735 & 21.2276 & 20.8927 & 19.9440 \\ 18.4736 & 19.9608 & 20.8927 & 21.1853 & 20.8230 \\ 16.6762 & 18.5264 & 19.9440 & 20.8230 & 21.0965 \end{bmatrix}$$

The SVD yielded

$$U_2 = \begin{bmatrix} 0.8905 & 0.4123 & 0.0935 & -0.1344 & -0.1008 \\ 0.3705 & -0.4959 & -0.6851 & 0.2227 & 0.3128 \\ 0.2259 & -0.4936 & 0.5293 & 0.5778 & -0.3023 \\ 0.1280 & -0.4076 & 0.4613 & -0.5170 & 0.5808 \\ 0.0473 & -0.4175 & -0.1701 & -0.5755 & -0.6807 \end{bmatrix}$$

$$\Sigma_2 = \begin{bmatrix} 9.7032 & 0 & 0 & 0 & 0 \\ 0 & 1.7256 & 0 & 0 & 0 \\ 0 & 0 & 0.0533 & 0 & 0 \\ 0 & 0 & 0 & 0.0158 & 0 \\ 0 & 0 & 0 & 0 & 0.0117 \end{bmatrix}$$

$$V_2 = \begin{bmatrix} 0.4557 & -0.6682 & 0.5738 & -0.1019 & 0.0784 \\ 0.4670 & -0.2754 & -0.5643 & 0.6055 & -0.1447 \\ 0.4617 & 0.0760 & -0.4483 & -0.6411 & 0.4112 \\ 0.4410 & 0.3630 & 0.1371 & -0.2325 & -0.7752 \\ 0.4081 & 0.5832 & 0.3640 & 0.3974 & 0.4504 \end{bmatrix}$$

The resulting values of AIC(k) were

$$\begin{aligned} AIC(1) &= 0.3771 \\ AIC(2) &= -\,2.7253 \\ AIC(3) &= -\,2.7595 \longleftarrow \\ AIC(4) &= -\,2.6753 \\ AIC(5) &= -\,2.6253 \end{aligned}$$

so that $k_1^* = 3$ was selected. The resulting estimated model was

$$y(t) = 2.3931\, y(t-1) - 1.977\, y(t-2) + 0.7421\, y(t-3)$$

which is much closer to the actual model (6.12) than the estimate based on half as many data points.

The ΔAIC criterion was then applied to test for a model change. Using (6.13) with $k^* = 2$, $k_1^* = 3$, $S(k^*) = 0.0811$, $S_1(k_1^*) = 0.0564$, $\tilde{S}(k^*) = 0.1119$, we get

$$\Delta AIC = 0.3632 + 0.3801 - 0.15 = 0.5933 \qquad (6.14)$$

which yields a "change" decision. By comparing (6.14) to (6.13) we note several things. The first term, which is $\log S(k^*) - \log S_1(k_1^*)$ now more strongly indicates a change, due to better model fit. The second term, which is the effect of modeling error on the measured error covariances, also more strongly indicates a change due to increased data length, which produces a more accurate estimate of the true error covariance during the testing interval, using the "no change" hypothesis. Finally, the last term more strongly indicates a change, since the bias for a "no change" decision is reduced due to increased data length. Thus we see that all three terms in the ΔAIC criterion contribute to the final decision, and each one is of importance in achieving an accurate decision.

7. DETECTION OF SLOW MODEL CHANGES

We now consider the detection of slow, essentially continuous, model changes. What we wish to achieve in this case is an appropriate data length over which to fit models. If the data length is too short, then there will be a tendency to overfit the model to the noisy data, leading to larger prediction errors. If the data length is too long, then the effects of parameter variations will begin to dominate the prediction errors.

In order to generate an appropriate measure by which to trade off these two characteristics, we again use the AIC criterion, but in a different way. Assume we have data over a time interval $I = \{1,2, \ldots, n\}$ and suppose we divide this interval into subintervals of length W:

$$I_1 = \{1,2, \ldots, W\}$$

$$I_2 = \{W+1, W+2, \ldots, 2W\},$$

etc.

Then an appropriate measure for data length determination is the average per sample entropy. In terms of the AIC criterion we define

$$\overline{AIC}_W = \frac{1}{N_W} \sum_i \frac{1}{W} AIC_p(k_i^*)$$

where $AIC_p(k_i^*)$ is the minimum <u>prediction</u> AIC for the $\underline{i^{th}}$ interval I_i, k_i^* is the optimal model order for the $\underline{i^{th}}$ interval, and N_W is the number of intervals of W over the whole data interval I. The prediction AIC uses the forward prediction error variance (cf eq. (5.9)) rather than the fit error

variance, since we are interested in the error over the next interval, not the one over which the model was fitted. This has the effect of increasing the penalty on the number of parameters in the AIC criterion.

The form of AIC_p is

$$AIC_p(k_i^*) = AIC(k_i^*) + M(k_i^*)$$

Experimental Results

In order to test this criterion as the basis for data length selection, we considered a second-order AR model

$$y(t) = a_1(t)\, y(t-1) + a_2(t)\, y(t-2) + n(t)$$

where $n(t)$ was zero-mean white gaussian noise with unit variance. The time varying coefficients were selected so that the two system poles were on the unit circle in the z-plane. This yields: $a_1(t) = 2 \cos \Theta(t)$, $a_2 = -1$ where the roots are: $\cos \Theta(t) + i \sin \Theta(t)$ and $\cos \Theta(t) - i \sin \Theta(t)$. The time-variation of $\Theta(t)$ was selected as $\Theta(t) = \Theta(0) + 2 \pi f t$, where f is a selected frequency. Two values of f (.0001, .001) were used in the experiments. Total data length was 1000 time points. The results for Case 1 ($f = 0.001$) are shown in Table 7.1, using $\Theta(0) = 0.2$. The result is that the optimal indicated data length is 10-12 samples and corresponds to the case in which the average coefficient change over the fit window is in the range of 0.80 - 0.96. Over the entire data length of 1000 samples, the value of a_1 starts at 1.64, deceases to - 1.90 at t = 400 and then increases to 1.90 at t = 1000. Thus, the average coefficient change over the optimum data length is generally more than 40% of the coefficient value.

Table 2 shows the results for Case 2 ($f = 0.0001$) in which the optimal data length is found to be 30 samples. This is, of course, increased over that of Case 1 since the coefficients vary much less rapidly - on the order of 0.019, on the average. Note that the rms prediction error σ_e evaluated over the fit set generally decreases monotonically with data length and cannot be used as a selection criterion.

Table 7.1
Data Length Estimation for Case 1 (f = 0.001)

W	$\overline{AIC_W}$	$\Delta\overline{AI}$	σ_e	σ_{a1}	σ_{a2}
100	0.521	0.768	1.817	0.163	0.086
20	-1.364	0.159	0.603	0.118	0.109
15	-1.479	0.119	0.508	0.141	0.110
12	-1.557	0.096	0.520	0.116	0.108
10	-1.557	0.080	0.487	0.152	0.139
8	-1.512	0.064	0.469	0.133	0.125
5	-1.265	0.0399	0.389	0.193	0.200

W = data length of window

$\overline{AIC_W}$ = average AIC

$\Delta\overline{AI}$ = average change of AI over window

σ_e = rms one-step-ahead prediction error over fit window

σ_{a1} = rms error in a1 estimate

σ_{a2} = rms error in a2 estimate

Table 7.2
Data Length Estimation for Case 2 (f = 0.0001)

W	$\overline{AIC}_W$	$\Delta\overline{AI}$	σ_e	σ_{a1}	σ_{a2}
100	-1.959	0.061	0.519	0.016	0.018
50	-2.131	0.031	0.486	0.032	0.031
40	-2.152	0.025	0.464	0.055	0.049
30	-2.153	0.019	0.437	0.085	0.068
20	-2.052	0.012	0.462	0.507	0.046
10	-1.835	0.0061	0.455	0.065	0.062

REFERENCES

Aitchison J. and I.R. Dunsmore (1975). Statistical Prediction Analysis, Cambridge University Press.

Akaike, H. (1976). "Canonical Correlation Analysis of Time Series and the Use of an Information Criterion." System Identification: Advances and Case Studies, R.K. Mehra and D.G. Lainiotis, eds., New York: Academic Press, pp. 27-96.

Akaike, H. (1975). "Markovian Representation of Stochastic Processes by Canonical Variables." SIAM J. Contr., Vol 13, pp. 1620-173

Akaike, H. (1974a). "Stochastic Theory of Minimal Realization." IEEE Trans. Automat. Contr., Vol. 19, pp. 667-674.

Akaike, H. (1974b). "A New Look at Statistical Model Identification." IEEE Automatic Control, Vol 19, pp. 667-674.

Akaike, H., (1973). "Information theory and an extension of the maximum likelihood principle." In 2nd International Symposium on Information Theory., Eds. B.N. Petrov and F. Csaki, pp. 267-281. Budapest: Akademiai Kiado.

Anderson, T.W. (1971). The Statistical Analysis of Time Series. New York: Wiley.

Astrom, K.J. (1973). "On self-tuning regulators." Automatica, Vol 9, pp. 185.

Astrom, K.J. (1983). "Theory and Applications of Adaptive Control-A Survey." Automatica, Vol 19, pp. 471-86.

Astrom, K.J., Borisson, L. Ljung and B. Wittenmark, (1977). "Theory and applications of self-tuning regulators." Automatics, Vol 13, pp. 457.

Astrom, K.J. and B. Wittenmark (1973). "On self-tuning regulators", Automatica, Vol 9, pp 185.

Bhansali, R.J., and Downham, D.Y. (1977), "Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion." Biometrika, Vol 64, pp. 547-71.

Box, G.E.P. and G.M. Jenkins (1970), Time Series Analysis Forecasting and Control, San Francisco: Holden-Da.

Cox, D.R. and Hinkley, D.V. (1974). Theoretical Statistics. London: Chapman and Hall.

Desai, U.B., and D. Pal (1982), "A Realization Approach to Stochastic Model Reduction and Balanced Stochastic Realization," Proc. 21st Conf. on Decision and Control, Orlando, pp. 1105-12.

Elliott, H. and W.A. Wolovich (1984). "Parameterization Issues in Multivariable Adaptive Control." Automatica, Vol 20, pp. 533-45.

Evans, R.J. and R.E. Betz (1982). "New Results and Applications of Adaptive Control to Classes of Nonlinear Systems." Proc. Workshop on Adaptive Control. Florence, Italy.

Faurre, P.F. (1976). "Stochastic Realization Algorithm." System Identification: Advances and Case Studies, R.K. Mehra and D.G. Lainiotis, eds, New York: Academic Press, pp. 1-25.

Friedlander, B. (1982). "Lattice Filters for Adaptive Processing," Proc. IEEE, Vol 70, pp. 829-67.

Friedlander, B. (1982). "Lattice Methods for Spectral Estimation," Proc. IEEE, Vol 70, pp 990-1017.

Friedlander, B. (1983). "Lattic Implemantation of Some Recursive Parameter-Estimation Algorithms," Int. J. Control, Vol. 37, pp. 661-684.

Gerhardt, L.A. (1978). "A Comparison of Spectrum Estimation Techniques Using Common Data Sets," Proceeding of the RADC Spectrum Estimation Workshop, held May 24-6, 1978, at Rome Air Force Base, Rome, NY. Technical Report No. AD-A054650,Defense Technical Information Center, Alexandria, VA.

Gevers, M. and V. Wertz (1982). "On the Problem of Structure Seletion for the Identification of Stationary Stochastic Processes," Papers of the IFAC Symp. on Identification and System Parameter Estimation, G. Bekey and G. Saridis (eds.), Wash. D.C.: Mcgregor-Werner, pp. 387-92.

Goldstein, J.D., and W.E. Larimore, (1980). "Applications of Kalman Filtering and Maximum Likelihood Parameter Identification to Hydrologic Forecasting." The Analytic Sciences Corporation, Report No. TR-1480-1, March 1980. Available as Report AD-A113347 through Defense Technical Information Center, Alexandria, VA 23314.

Golub, G.H. (1969). Matrix Decompositions and Statistical Calculations. Statistical Computation, R.C. Milton and J.A. Nelder, eds., New York: Academic Press, pp. 365-379.

Goodrich, R., W.E. Larimore and R.K. Mehra (1983). "New Results in State Space Forecasting." International Symposium on Forecasting, Philadelphia, PA, June 5-8.

Granger, C.W.J. and G. McCollister (1979). "Comparison of Forecasts of Selected Series by Adaptive." Box-Jenkins and State Space Methods, ORSA/TIMS, Los Angeles, California.

Hagglund, T. (1983). "New Estimation Techniques for Adaptive Control." Report CODEN:LUTFD2/(TFRT-1025)/1-120/(1983), Department of Automatic Control, Lund Institute of Technology. Doctoral Dissertation.

Hart, P.E. (1971). "Entropy and Other Measures of Concentration," J. Roy Statist. Soc., A, Vol 134, pp. 73-85.

Honig, M.L., and D.G. Messerschmitt (1984). Adaptive Filters: Structures, Algorithms, and Applications. Boston: Kluwer Academic Publishers.

Hotelling, H. (1936). "Relations between Two Sets of Variates." Biometrika, Vol 28, pp. 321-377.

Irving, E. (1979). "New Development in improving power network stability with adaptive control." Proc. Workshop on Applications of Adaptive Control. Yale University, New Haven.

Isermann, R. (1984). "Process Fault Detection Based on Modeling and Estimation Methods - A Survey," Automatica, Vol 20, pp. 387-404.

Jeffreys, H. (1961). Theory of Probability, Clarendon Press.

Kendall, M.G. (1973). "Entropy, Probability and Information," International Statistical Review, Vol 41, pp. 59-68.

Kullback, S. (1959), Information Theory and Statistics, Dover.

Kullback, S. and R.A. Leibler (1951), "On Information and Sufficiency," Ann. Math. Statistics, 22, pp. 79-86.

Kung, S.Y. and D.W. Llin (1981). "Optimal Hankel-Norm Model Reductions: Multivariable Systems", Trans. Auto. Control, Vol 26, pp. 832-852.

Larimore, W.E. and R.K. Mehra (1984), "Technical Assessment of Adaptive Flutter Suppression Research," Air Force Wright Aeronautical Lab Report No-AFWAL-TR-84-3052, SSI.

Larimore, W.E., S. Mahmood and R.K. Mehra (1983). "Adaptive Model Algorithms Control." Proc. IFAC Workshop on Adaptive Systems in Control and Signal Processing, San Francisco, CA, June 1983.

Larimore, W.E. (1983a). "Predictive inference, sufficiency, entropy, and an asymptotic likelihood principal." Biometrika, Vol 70, pp. 175-81.

Larimore, W.E. (1983b). "Systems Identification, Reduced-Order Filtering and Modeling Via Canonical Variate Analysis." Proc. 1983 American Control Conference, H.S. Rao and T. Dorato, eds., New York: IEEE. pp. 445-51.

Larimore, W.E. (1981a). "Small sample methods for maximum likelihood identification of dynamical processes." Applied Time Series Analysis, Proceedings of the Fifth International Time Series Meeting. Houston, Texas, August 1981. Amsterdam, North Holland, pp. 167-174.

Larimore, W.E. (1981b). "Recursive maximum likelihood and related algorithms for parameter identification of dynamical processes." Proceedings of the 20th IEEE Conference on Decision and Control, Vol 1, pp. 50-55, San Diego, California, December 1981.

Larimore, W.E. (1977). "Nontested Tests on Model Structure." Proceedings Joint Automatic Control Conf. (San Francisco, CA). New York: IEEE, pp. 686-690.

Larimore, W.E. (1977b). "Statistical inference on stationary random fields." Proc. IEEE, Vol 65, pp. 961-70.

Loy, X.C., A.S. Willsky, and G.C. Verghese (1983). "Failure with Uncertain Models," Proc. Amer. Control Conf. San Francisco, California.

Ljung, L. and T. Soderstrom (1983). "Theory and Practice of Recursive Identification." Cambridge: MIT Press.

Ljung, L. (1979). "Asymptotic Behavior of the Extended Kalman Filter as a Parameter Estimator for Linear Systems." IEEE Trans. Auto. Control, Vol 24, pp. 36--50.

Ljung, L., I. Gustavsson and T. Soderstrom (1974). "Identification of Linear Multivariable Systems Operating Under Linear Feedback Control." IEEE Trans. Auto. Control, AC-19, pp. 836-840.

Mehra, R.K. (1982). "Identification in Control and Econometrics." Current Development in the Interface: Economics, Econometrics, Mathematics. M. Hazewinkel and A.H.G. Rinnooy Kan, Eds., Dordrecht, Holland: Reidel, pp. 261-288.

Mehra, R.K., (1981). Choice of Input Signals. Trends and Progress in System Identification, P. Eykhoff, ed., IFAC, Oxford: Pergamon Press.

Mehra, R.K. (1978). A Survey of Time-Series Modeling and Forecasting Methodology. Modeling, Identification and Control in Environmental Systems, E. Vansteenkiste, ed., North Holland Publishing Co.

Mehra, R.K. and A. Cameron (1980). "Handbook on Business and Economic Forecasting for Single and Multiple Time Series." Scientific Systems, Inc., Notes for the Institute of Professional Education Seminar.

Mehra, R.K. and A. Cameron (1976). "A Multidimensional Identification and Forecasting Technique Using State Space Models." ORSA/TIMS Conf. Miami, FL, November 1976.

Mehra, R.K. and J.S. Tyler (1973). "Case Studies in Aircraft Parameter Identification." Proc. 3rd IFAC Conf. on Identification and System Parameter Estimation, P. Eykhoff, ed., Oxford: Pergamon Press, 117-144.

Mehra, R.K. and J. Peschon (1971). "An innovations approach to fault detection and diagnosis in dynamic systems." Automatica, Vol 7, pp. 657.

Murray, G.D. (1977). "A note on the estimation of probability density functions." Biometrika, Vol 64, pp. 150-2.

Murray, G.D. (1979). "The estimation of multivariate normal density functions using incomplete data." Biometrika, Vol 66, pp. 375-80.

Peloubet, R.P., Jr., R.L. Haller and R.M. Bolding, (1980). "F-16 Flutter Suppression System Investigation", presented at the AIAA/ASME/ASCE/AHS 21st Structures, Structural Dynamics and Materials Conference, Seattle, Washington, May 1980.

Schwartz, G. (1978). "Estimating the Dimension of a Model," Ann. Statistics, Vol 6, 2, pp. 461-4.

Shibata, R., (1981a). "An optimal autoregressive spectral estimate." Ann. Statist. Vol 9, pp. 300-6.

Shibata, R., (1981b). "An optimal selection of regression varibles." Biometrika, Vol 68, pp. 45-54.

Soderstrom, T., L. Ljung and I. Gustavsson, (1978). "A Theoretical Analysis of Recursive Identification Algorithms." Automatica, Vol 14, pp. 231-244.

Wald, A., (1943). "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large." Trans. Amer. Math. Soc., Vol 54.

Wellstead, P.E. and S.P. Sanoff, (1981). "Extended self-tuning algorithm." Intl. J. Control. Vol 34, pp. 433-455.

Whittle, P., (1953). "The Analysis of Multiple Stationary Time Series," J.R. Statist. Soc B., Vol 15, pp. 125-39.

Willsky, A.S., (1980). "Failure Detection in dynamic systems." AGARD, No. 109.

Willsky, A.S., (1976). "A Survey of Design Methods for Failure Detection Systems," Automatica, Vol 12, pp. 601-611.

Willsky, A.S., (1976). "A Survey of Design Methods for Failure Detection in Dynamic Systems." Automatica, Vol 12, 601-11. Decision and Control, December 8-10, Orlando, FL, pp. 148-151.

Willsky, A.S. and H.L. Jones (1974). "A Generalized likelihood ratio approach to state estimation in linear systems subject to abrupt changes." Proc. IEEE Conf. on Decision and Control. Phoenix, Arizona.

APPENDIX

MAXIMUM LIKELIHOOD ESTIMATION

We present here some basic background on maximum likelihood estimation, which is used throughout this report.

The likelihood function for a sample $x_1, x_2, \ldots x_n$ parametrized by a parameter Θ is

$$L = \prod_{i=1}^{n} p(x_i \mid \Theta) \tag{A.1}$$

Assume the x_i are drawn independently from the true distribution $p(x \mid \Theta_0)$. Then L is the joint distribution function of $x_1, x_2, \ldots, x_n$ and

$$\int\ldots\int L \, dx, \ldots dx_n = 1 \tag{A.2}$$

Differentiating wrt Θ:

$$\int\ldots\int \left(\frac{\partial L}{\partial \Theta}\right) dx, \ldots dx_n = 0 \; ; \; \frac{\partial L}{\partial \Theta} = \text{row vector}$$

or

$$\int\ldots\int \left(\frac{\partial \log L}{\partial \Theta}\right) L \, dx, \ldots dx_n = 0$$

or

$$E\left(\frac{\partial \log L}{\partial \Theta}\right) = 0 \tag{A.3}$$

Differentiating again

$$\int \cdots \int \left(\frac{\partial^2 L}{\partial \Theta^2}\right) dx_1 \ldots dx_n = 0$$

Now

$$\frac{\partial^2 \log L}{\partial \Theta^2} = -\frac{1}{L^2}\left(\frac{\partial L}{\partial \Theta}\right)^T \left(\frac{\partial L}{\partial \Theta}\right) + \frac{1}{L}\frac{\partial^2 L}{\partial \Theta^2}$$

so that

$$\frac{\partial^2 \log L}{\partial \Theta^2} = -\left(\frac{\partial \log L}{\partial \Theta}\right)^T \left(\frac{\partial \log L}{\partial \Theta}\right) + \frac{1}{L}\frac{\partial^2 L}{\partial \Theta^2}$$

Thus

$$E\left[\frac{\partial^2 \log L}{\partial \Theta^2}\right] = -E\left[\left(\frac{\partial \log L}{\partial \Theta}\right)^T \left(\frac{\partial \log L}{\partial \Theta}\right)\right]$$

Now

$$\left.\frac{\partial \log L}{\partial \Theta}\right|_{\Theta_0} = \left.\frac{\partial \log L}{\partial \Theta}\right|_{\hat{\Theta}} + (\Theta_0 - \hat{\Theta})^T \left.\frac{\partial^2 \log L}{\partial \Theta^2}\right|_{\hat{\Theta}}$$

where $\hat{\Theta}$ is the maximum likelihood estimate which satisfies

$$\left.\frac{\partial \log L}{\partial \Theta}\right|_{\hat{\Theta}} = 0$$

and Θ_0 is the true parameter value.

Thus

$$\left.\frac{\partial \log L}{\partial \Theta}\right|_{\Theta_0} = (\Theta_0 - \hat{\Theta})^T \left.\frac{\partial^2 \log L}{\partial \Theta^2}\right|_{\hat{\Theta}} \tag{A.4}$$

Now define the covariance matrix

$$C = E\left[\left(\frac{\partial \log L}{\partial \Theta}\right)\Bigg|_{\hat{\Theta}}^{T} \left(\frac{\partial \log L}{\partial \Theta}\right)\Bigg|_{\hat{\Theta}}\right] = - E\left[\left.\frac{\partial^2 \log L}{\partial \Theta^2}\right|_{\hat{\Theta}}\right] \tag{A.5}$$

and factor C as $C = W W^T$.

Write (A.4) as

$$\left.\frac{\partial \log L}{\partial \Theta}\right|_{\Theta_0} W^{-T} = (\Theta_0 - \hat{\Theta})^T \left.\frac{\partial^2 \log L}{\partial \Theta^2}\right|_{\hat{\Theta}} W^{-T}$$

The right hand side is approximated as

$$(\Theta_0 - \hat{\Theta})^T C W^{-T} = (\Theta_0 - \hat{\Theta})^T W$$

The left hand side is a normalized gaussian variate since

$$E\left\{\left[\left(\frac{\partial \log L}{\partial \Theta}\right)\Bigg|_{\hat{\Theta}} W^{-T}\right]^T \left[\left.\frac{\partial \log L}{\partial \Theta}\right|_{\hat{\Theta}} W^{-T}\right]\right\} = I$$

Thus, the right hand side is also a normalized gaussian variate and

$$E\left\{[(\Theta_0 - \hat{\Theta})^T W]^T [(\Theta_0 - \hat{\Theta})^T W]\right\} = I$$

which yields

$$E\,[(\Theta_0 - \hat{\Theta})\,(\Theta_0 - \hat{\Theta})^T] = C^{-1} \qquad \text{(A.6)}$$

C is the Fisher information matrix, which is the inverse of the covariance matrix of the parameter estimation errors.

END
DATE
FILMED
5-88
DTIC